4-2021

Classification of Primary versus Metastatic Pancreatic Tumor Cells Using Multiple Biomarkers and Whole Slide Imaging

Poupack Pooshang Baghery
University of Nebraska-Lincoln, ppooshang-baghery@huskers.unl.edu

CLASSIFICATION OF PRIMARY VERSUS METASTATIC PANCREATIC TUMOR CELLS USING MULTIPLE BIOMARKERS AND WHOLE SLIDE IMAGING

by

Poupack Pooshang Baghery

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Electrical Engineering

Under the Supervision of Professor Khalid Sayood

Lincoln, Nebraska

April 2021

CLASSIFICATION OF PRIMARY VERSUS METASTATIC PANCREATIC TUMOR CELLS USING MULTIPLE BIOMARKERS AND WHOLE SLIDE IMAGING

Poupack Pooshang Baghery, M.S.

University of Nebraska, 2021

Advisor: Khalid Sayood

Pancreatic cancer is a challenging cancer with a high mortality rate and a 5-year survival rate between 2% to 9%. The role of biomarkers is crucial in cancer prognosis, diagnosis, and predicting the possible responses to a specific therapy. The Discovery and development of various types of biomarkers have been studied intensively in the hope of determining the best treatment approaches, better management, and possibly cure of this deadly cancer. However, metastasis, responsible for about 90% of the deaths from cancer, is still poorly understood. A few research that have investigated the expression of a particular biomarker or a panel of biomarkers in the primary and secondary (metastatic) tumor demonstrates that the expression of different biomarkers in the primary and secondary tumor sites is not necessarily the same, even though the primary and metastatic tumor cells are originated from the same organ. In this project, we aim to design a classifier to distinguish between primary and secondary tumor cells based on their uptake of different biomarkers, using immunofluorescence whole slide imaging. For this purpose, we first register consecutive images of the same slide together to be able to locate multiple biomarkers that belong to a cell and later we design our classifier based on vectors that show the presence or absence of multiple antibodies in addition to the amount of that antibody in a tumor cell.

Acknowledgments

I would like to express my deepest appreciation to my advisor, Dr. Khalid Sayood, for his treasured support, invaluable knowledge, patience, and kindness during the course of my M.S degree. He has truly been an amazing mentor and has provided the guidance I have needed throughout my research. I would also like to thank Dr. Walter Scott Campbell and Dr. Heather Jensen Smith for their precious time, support and assistance in the progression of this research project. I wish to thank my thesis committee members, Dr. Walter Scott Campbell, Dr. Hasan Otu and Dr. Andrew Harms, for all their help, insightful comments and suggestions.

I would like to thank my friends and lab mates, Amir Salar, Dicle, Sree, Bridget, Joel and Brittany for a warm friendly atmosphere they created in our lab.

Finally, I am especially grateful for my wonderful family for their unconditional love, support, and encouragement over the years.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1. Pancreatic cancer

Pancreatic cancer initiates in the pancreas, a body organ located behind the stomach and next to the small intestine, Fig 1. This organ, which consists of three parts: head, body, and tail, has two kinds of cells: endocrine cells and exocrine cells. Endocrine cells secrete hormones such as insulin to regulate blood sugar. Exocrine cells release enzymes to help with food digestion. The different kinds of pancreatic cancer are divided into two main groups; exocrine tumors and neuroendocrine tumors.

Pancreatic Ductal Adenocarcinoma (PDAC), an exocrine tumor, is one of the most common types of pancreatic cancer that occurs in the pancreatic ducts. Due to the location of the pancreas in the belly, pancreatic tumors are not usually felt by pressing the belly. The symptoms appear when cancer has spread to other body organs, which explains why pancreatic cancer is rarely diagnosed at the stage when it could be cured [1].

Despite all advances in cancer treatment, this malignancy still remains one of the deadliest cancers. Its 5-year survival rate ranges from 2% to 9 % in the United States [2], and approximately 7% of all cancer deaths come from this type of cancer [3]. Although the main causes of this type of cancer are not yet identified, factors such as smoking, obesity, and specific gene mutations such as KRAS mutation may affect the chance of getting this disease.

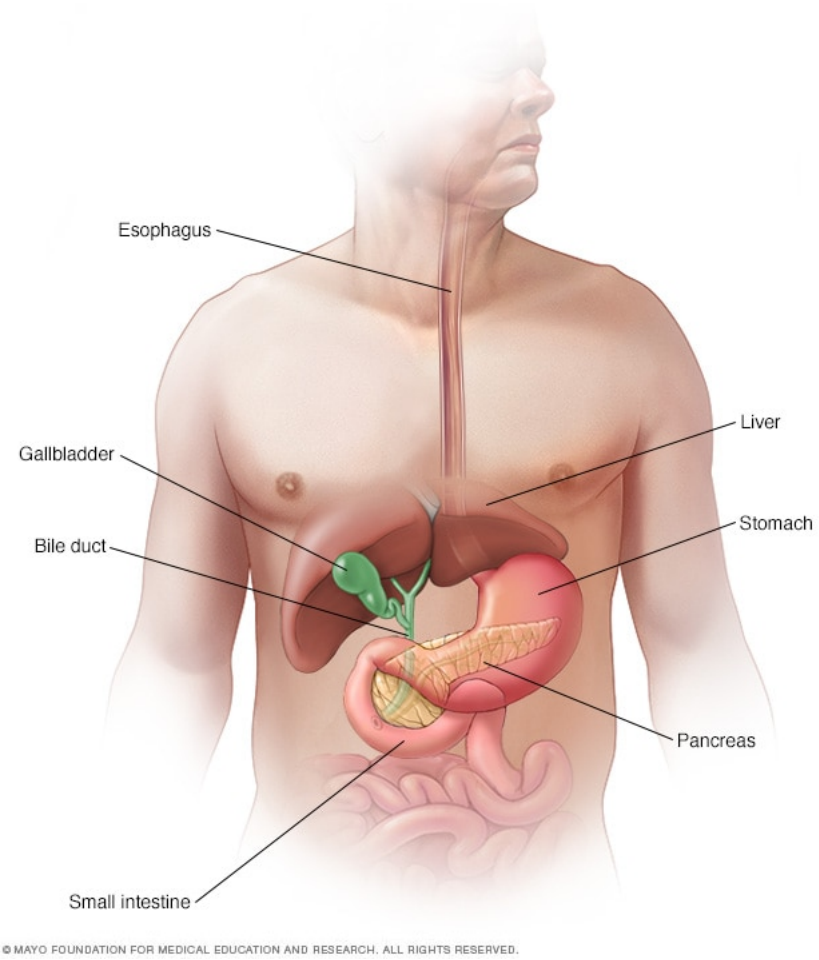

Fig 1. Location of the pancreas in the body (Image courtesy Mayo Clinic)

Although this challenging cancer is not curable and has an abysmal survival rate, fast-paced cancer studies in recent years have deepened our understanding of the biology of pancreatic cancer which in turn has affected the management of this cancer in different aspects such as early detection, medical therapy, and drug development. According to the World Health Organization (WHO), a biomarker is defined as "any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease."[4] . The discovery and development of minimally invasive robust biomarkers with high sensitivity and specificity is one of the most

important steps for moving toward personalized medicine and avoid unnecessary treatments with the high cost and adverse side effects on the patients [5]. Also, different imaging modalities in combination with biomarkers have provided ample opportunities for improving cancer detection and treatment and have made designing patient-specific therapies for different types of cancer more feasible while the biomarker assessment indicates the presence of any disease, imaging techniques can facilitate this task. Also, they can be harnessed to locate the tumor and identify the aggressiveness of the tumor [6].Whole slide imaging is one of the most recent imaging modalities with novel technology, which has several clinical and non-clinical applications that can be used to test for the presence of any cancerous antibody.

1.2. Problem statement and Organization of this Thesis

The ultimate goal of this project from the outset was to develop an algorithm that would be able to classify pancreatic cancer cells into primary or metastatic (secondary) categories based on the differential uptake of antibodies as a tumor marker. Sets of antibodies with fluorescent labels are applied to the tissue and imaged. The tissue is washed between applications. Since a glass slide is manually located inside the tray of a whole slide image scanner, each rescan of the tissue results in an image with the tissue at a slightly different orientation and therefore different coordinates for collocated pixels, consequently, coordination of the points will change. This problem is resolved by registering the consecutive images of a slide together. In order to analyze the presence or absence of all antibodies in consecutive images of a tissue, prior to ascertaining which cells are taking up which antibodies, the different scans must be registered. Once the scans are registered, the location of the cells taking up different antibodies can be specified and used

in addition to each antibody's expression associated with the cell. Together, these form the components of a feature vector, all vectors create a big matrix which will be the input to machine learning algorithms developed for classifying the cells. In this research, after an introduction about pancreatic cancer in chapter one, we will give an overview of the problem and explain how we will be dealing with this problem. In chapter two, we will describe the digital pathology or whole slide imaging system, its applications, benefits, and drawbacks. Later we broadly talk about biomarkers and tumor environment in pancreatic cancer and explain how the use of biomarkers is important in the prognosis, diagnosis, and prediction of different therapeutic approaches in pancreatic cancer and how intensive research in this area is bringing hope for better management and even cure of this deadly disease.

In chapter 3, we will be working with thumbnail brightfield and immunofluorescence images. First, we register two consecutive brightfield thumbnail images based on finding fiducial markers, and in the second part we will be registering immunofluorescence images of the tissue samples.

In the evaluation of differences between normal and cancer cells - the details of which are outside the scope of this project - usually morphology differences are analyzed. Cancerous cells are different from normal cells in both shape and size. Figure 2 shows some of the morphological differences between normal and cancerous cells. These differences give the opportunity to design a classifier based on a single or a panel of features including area, axes lengths, eccentricity, perimeter, circularity, and other morphological characteristics. Most of the classifiers are designed based on normal versus cancerous cells. However, in chapter 4, we implement two types of classifiers to investigate

the differences between primary tumor and metastatic tumor cells; The first type is a morphology-based classifier to analyze if primary tumor and metastatic tumor cells are distinguishable morphologically. Finally, the second type of the classifier is designed to assess how biomarker uptakes of primary tumor cells differ from their metastatic tumor counterparts.

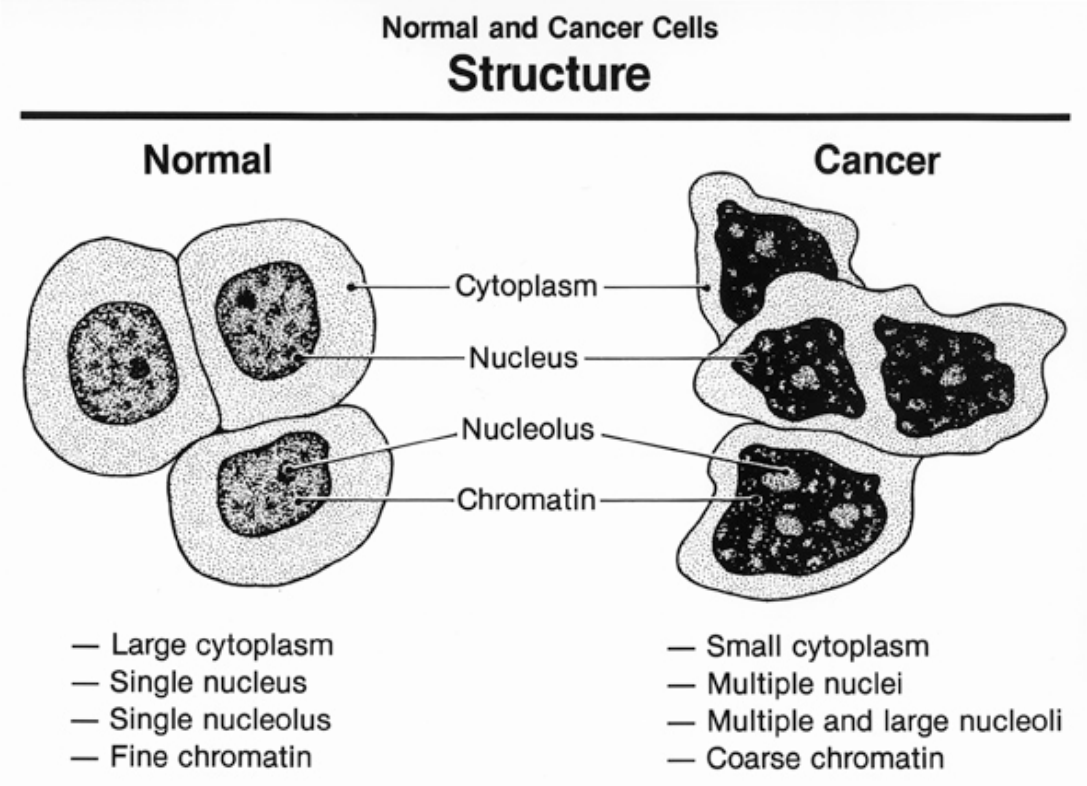

Fig 2. Normal cells versus cancer cells morphological characteristics [7]

The proposal of this project was based on immunohistochemical brightfield whole slide images of pancreatic cancer tissues, however, we adjusted the research to the data sets

that we received. Those data sets were received sequentially over the course of this project. First, we found the fiducial markers inside the thumbnail images of a whole slide image and registered two thumbnail images together based on these markers. Later, we received two non-consecutive large brightfield whole slide images, and we started to work on those. Shortly after, we received several small size brightfield images in different proprietary formats and worked to change the format into .tiff or .jpeg to be able to process them. We experienced a loss of information while converting to .jpeg or .tiff, and this is unacceptable for the accuracy standards of pathology. Finally, we received immunofluorescence whole slide images in small sizes (1017 x 1920 pixels) of primary and metastatic tumors. Only for two patients did we have both primary and metastatic information. The images for the remaining patients are either primary or metastatic without any ground truth. The public datasets are mostly brightfield images of different types of cancer, and our aim was to classify pancreatic cancer cells based on either morphology characteristics or uptaking antibodies in immunofluorescence whole slide images. Besides, pancreatic cancer is difficult for pathologists because “acini are cut obliquely, making it difficult to discern their characteristics shape” [8], therefore images of other types of cancer are an insufficient substitute for helping solve the challenge of pancreatic cancer.

Chapter 2
Background

Before we start describing methods of registration of immunofluorescence whole slide imaging and apply them to see the results of the registration and to locate multiple biomarkers in chapter 3 and later classification of primary versus metastatic cancer cells in pancreatic cancer in chapter 4, we introduce the whole slide imaging system (WSI), the notion of biomarker and tumor microenvironment in this chapter. Also, based on reviewing the existing literature we provide the background behind these concepts, explain how tumor environment contributes in tumor progression and metastasis, discuss potential biological markers in pancreatic cancer and how they can be promising in prognosis, diagnosis or prediction of responses to different therapeutic approaches.

The quality of the microscope was enhanced since 1850, which paved the way for the first pathology practices to start. Since that time, pathologists have been using the traditional microscope as the gold standard for the diagnosis of cancer and other diseases. Nevertheless, advances in digital imaging and image processing have opened a window to move from traditional microscopy to virtual microscopy [9]. Figure 3 presents the evolution of pathology over time.

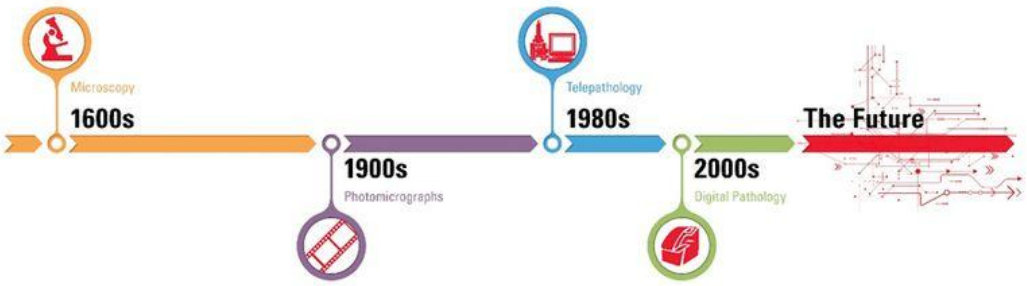

Fig 3. The evolution of pathology over time [10]

Wetzel and Gilbertson developed the first high-resolution whole-slide imaging (WSI) system in 1999 [11]. In digital pathology, or Whole slide imaging, the glass slides are scanned by a high-resolution scanning device under brightfield or fluorescent conditions, and a high-resolution digital image of the entire microscope slide is created [12]. The total scan time of a glass slide is less than one minute, with a resolution comparable to seeing glass slides under a traditional microscope. Scanning can occur at different magnifications. However, 20X magnification is acceptable for routine image analysis. The digitization process includes scanning, storage, editing, and display. Each whole slide imaging system consists of these components: light source, slide stage, objective lenses, and a high-resolution camera [12]. In brightfield microscopy, which is the most common type of microscopy, the light passing by the specimen is collected by the objective lens beneath the specimen [13]. In fluorescent microscopy, specific structures in the sample are labeled with fluorescent dyes named fluorophores. Only a few things in the tissue sample are labeled with these fluorophores and therefore light up. This enables practitioners to only see and focus on particular structures or objects instead of everything on the entire slide [14].

Between brightfield and immunofluorescence microscopy, which are both molecular examinations of a microscope slide to detect specific proteins within the tissue, brightfield is the most popular and preferred method for cancer diagnosis by pathologists. However, there is an increasing demand for utilizing immunofluorescence in multiplex biomarker detection because it can detect more than one biomarker per slide. Its counterpart, brightfield microscopy, is performed to detect only a single biomarker per slide. This might disrupt the tissue since it is being dyed and washed out several times [15]. In addition to its ability to detect multiple biomarkers, one of the most important advantages of using fluorescence microscopy is that intensity is linear with the amount of proteins. However, since low abundance proteins have weak signals, it is hard sometimes to pull out these signals from underneath the bright signals [16]. Fig 4 represents a whole slide image sample.

In digital pathology all the information, including slides and data can be examined, managed, and shared in a digital environment. Whole Slide Imaging (WSI) increases the workflow efficiency, offers decision support tools, and creates a connected team environment that allows for the sharing of slides, team annotations, and markups remotely. Also, Whole Slide Imaging could lead to considerable enhancement in translational research and computer-aided diagnosis (CAD). Also, digitization gives the ability to measure multiple areas of interest, evaluate several different viewing angles, views with more accuracy, and gain new and better insights from analyzing a massive number of images. All of this makes training, remote consultation, and clinical review easier [17].

Despite all the benefits of WSI, the high cost of scanners along with the need for a huge amount of storage for the digital files are some of its drawbacks. In addition to these

issues, there is also the challenge of the rate of pathologists who are familiar with this technology being low at this point in time [18].

Fig 4. Omnyx whole slide imaging scanner and viewer [18]

For the first time in 2017, the United States food and drug administration (FDA) approved the first commercial whole slide imaging (WSI) system, allowing the Philips IntelliSite Pathology Solution (PIPS) which reviews and interprets digital surgical pathology slides which have been prepared from biopsied tissue to enter the marketplace [19].

After the increased interest in the most recent imaging modality, whole slide imaging, for diagnostic, educational, and research purposes, one crucial question emerged.

Researchers and practitioners continue to wonder if whole slide imaging is as accurate as the optical microscopy that uses conventional glass slides. This led to a broad range of research related to clinical validation and standardization of this kind of imaging.

In several studies about surgical pathology - the analysis of a removed tissue from a living patient during surgery for diagnosis and treatment plan - WSI and traditional light microscopy have been compared. These studies show that WSI is not inferior to light microscopy, with one important caveat: WSI should be used for diagnosis purposes only when the pathologists are formally trained on the equipment. WSI has also been compared to traditional glass slides in the realm of primary diagnosis in anatomic pathology. It is likewise non-inferior to its long-established counterpart in this dimension either [20]. A meta-analysis, comprehensive literature search, among several publications from 2013 to 2019, which in total examined 10410 histology samples, demonstrated that there was a 98.3 % concordance between digital pathology (DP) and light microscopy (LM). Therefore, there is an equivalency between DP and LM in routine diagnosis. Although the discrepancies should be studied deeply before this emerging technology will take over permanently [21]. The College of American Pathologists states that each laboratory that works with whole slide systems, should conduct its own validation studies to be able to use digital pathology for diagnostic purposes [22].

Research in cancer biology shows that the progression of cancer is not solely related to changes in the tumor cells, as changes in the tumor microenvironment play a critical role in tumor development and progression [23]. There are multiple exchanges between cancerous cells and their neighboring microenvironment, and for understanding the

initiation, development, and progression of cancers a comprehensive analysis of tumor microenvironment to understand how it affects tumor growth and metastasis is essential. its mutual are highly dependent on interactions between the cancerous and nonmalignant cells in the tumor microenvironment [24-25].

The concept of tumor microenvironment (TME) dates back to 1889 when Stephen Paget after examining the data of 735 women with breast cancer and noticed that metastasis did not happen by chance, in fact for cancer cells (the seed) to metastasize a favorable microenvironment (the soil) is needed [26]. Tumor microenvironment is not only important in metastasis but also the dynamic interactions of cancer cells with cellular and acellular components of tumor microenvironment affects tumor growth and progression.

Pancreatic cancer is one of the most dangerous types of cancers with a high mortality rate and a dismal prognosis. Most patients are diagnosed at the stage when the tumor is locally advanced or has metastasized to other organs and therefore is not resectable. The close incidence rate and mortality rate of this malignancy has fueled the research to look deeply beyond the cancer cells and to further investigate the tumor microenvironment and its vital role in cancer progression to find novel therapeutic approaches for the treatment of pancreatic cancer.

The pancreatic cancer microenvironment is made up of cancer cells, tumor stromal cells, immune cells like macrophages and extracellular components. The components that are responsible for the progression of this malignancy are mainly regulatory T cells (Tregs),

tumor-associated macrophages (TAM), myeloid derived suppressor cells (MDSCs) and pancreatic stellate cells (PSC) [25],[27].

Recent studies show that the tumor microenvironment of pancreatic cancer, including cancer-associated fibroblasts such as stellate cells, extracellular matrix, different kinds of immune cells, and cytokines released by these cells, participates in the control of proliferation, invasion, and metastasis, chemoresistance and immunotherapy of pancreatic cancer by close interactions with cancer cells [27]; The dynamic interaction between tumor cells and their surrounding tissue favors the survival of the cancerous cells in such a way that the cancer cells divide and grow out of control following oncogenic mutations and therefore elude anti-tumor immunity, while the tumor environment of pancreatic cancer ductal adenocarcinoma can impact local immune response [25], [27].

There are two major characteristics of the pancreatic cancer microenvironment: dense desmoplastic reaction which is referred to suffusive growth of condensed fibrous tissue around the tumor, existing in both primary and metastatic tumors and extensive immunosuppression; Dense fibrous tissue prevents the infiltration of immune cells in the tumor tissue, making the tumor tissue escape from the surveillance of the immune system.

The desmoplasia builds a barrier around the tumor cells and therefore creates a hypoxic microenvironment in which prevents the proper formation of blood vessels and limits the exposure to chemotherapy and in consequence leads to poor immune cell infiltration. In such a hypoxic environment, the oxygen consumption is increased, and oxygen supply is compromised. Also, immunosuppressive molecules and cells by changing the balance of immune effector cells create a unique immunosuppressive environment that

facilitates cancer cell proliferation, the evasion of immune surveillance via the direct inhibition of anti-tumor immunity or the induction of immunosuppressive cell proliferation and metastasis [28]–[32].

Such an environment with these characteristics makes pancreatic cancer resistant to different kinds of therapy such as chemotherapy, radiation therapy and immunotherapy and this, in turn, promotes metastasis [28] ,[33]. Therefore, novel approaches to understand how different components of the pancreatic cancer microenvironment contributes to cancer progression and metastasis provide better insight to develop more effective treatments. For instance, an important question to be investigated is whether the tumor microenvironment characteristics in primary organ differ from the secondary organ when the tumor is metastasized [31]. Up to date, several approaches and methods to treat pancreatic cancer have failed or had unsatisfactory results Immunotherapy has improved cancer treatment significantly in several malignancies; however, pancreatic cancer due to its unique complex microenvironment remains unresponsive to conventional immunotherapies. However, advances in several fields such as biology, genetic and immunology with emerging tools like immunophenotyping, fluorescence multiplex imaging will facilitate deep understanding of the tumor microenvironment and successful personalized therapies hopefully in the near future [30-31], [34].

The role of biomarkers is crucial in cancer screening, prognosis, diagnosis, and determining the best treatment approaches. Understanding the relationship between biomarkers and their clinical results is of great significance not only to increase treatment options for all diseases but also to understand the normal status of the human body. Several

types of research have been performed since 1980 to examine the use of biomarkers in extremely important diseases such as cancer, and the FDA has continued to promote the use of biomarkers in clinical studies. However, biomarker-driven research has not been easy due to a variety of factors: the relatively low number of patients or healthy individuals that can be tested, the lack of assessment of the practicality of a proposed method, the selection of an early-stage group of patients, the healthy control groups, and the non-specificity of molecular markers.

Cancer biomarkers are classified into three groups; prognostic biomarkers that can provide valuable information to patients and assist clinicians in adjusting their treatment strategies according to the aggressiveness of the disease, diagnostic biomarkers that refers to those markers assisting with the early detection of cancer and potentially curable stage and predicting biomarkers that can help to predict how a patient might respond to treatment and how to select different treatment protocols based on a biological rationale in the very early stages of cancer that have the potential to improve patient survival rates [35]–[38].

In pancreatic cancer, one of the major challenges in biomarker development is obtaining tumor tissue samples of adequate quality for analysis. Initial diagnosis of this cancer is usually performed with fine needle aspiration (FNA), most commonly by endoscopic ultrasound, and therefore fair tissue procurement is difficult to obtain. Taking these biopsies is expensive, uncomfortable, and might lead to clinical complications.

These data highlight the need for biomarkers that are highly specific and easily measurable by inexpensive sensitive techniques so that could improve the diagnosis and accuracy of staging at the time of disease presentation to better inform first-line therapy

[38], In addition, due to the low incidence of pancreatic cancer in the population, stratification of the patients should be that accurate so that only those patients who truly need that, continue to undergo further examinations by invasive and expensive modalities. Therefore, minimally invasive modalities involving biomarkers and imaging techniques that would facilitate the early detection of pancreatic cancer are highly needed [39]- [40].

This complex biology and heterogeneity of cancer makes it hard to diagnose and treat effectively. Studies to develop novel potential biomarkers for diagnostic, predictive, and prognostic purposes have been an area of extensive research lately with the hope of finding effective management for this challenging cancer; however, none of them were used in clinical trials [41]-[42]. Serum carbohydrate antigen (CA) 19-9 was discovered in 1979 [43], and it is the most validated diagnostic marker in pancreatic cancer with sensitivity and specificity of 79-81% and 82-90%, respectively, but it is not useful in screening due to its low sensitivity and specificity. Several other carbohydrate antigens such as CA 50, CA-125, etc., have also been investigated, but studies demonstrate that they are overall less sensitive than CA19-9 [44].

Among strategies to find predictive, prognostic, and diagnostic biomarkers for pancreatic cancer liquid-based biopsies to detect circulating tumor cells (CTCs), circulating free DNA (cfDNA), and extracellular vesicles (EVs) are promising markers for early detection and diagnosis of PC [45]. These biomarkers, along with methylated DNA and exosomes, can classify the patients with pancreatic cancer adenocarcinoma and predict their sensitivity to the therapeutic methods [39]. Several studies show that exosomes correlate with pancreatic cancer progression and metastasis, and due to the possibility of detecting exosomes in different body fluids, they are considered as suitable potential

biomarkers in PC [46] Circular EV- based biomarkers are highly sensitive with high positive predictive value and low false positive value and offer an excellent opportunity for screening of individuals in pancreatic cancer [45]. Also, miRNA is another biomarker that has gained attention lately to be used as a marker for early detection of PC. For pancreatic cancer mass screening, affordable, convenient, and efficient testing with high sensitivity and specificity close to 100 % that can be utilized effectively for all the population is required [36].

Metastasis is responsible for the majority of cancer deaths; however, this phase of cancer has remained poorly understood. Most of the literature deals with differentiating between cancer versus normal situation in an organ [47]. A few studies have assessed the expression of particular biomarkers in primary versus metastatic tumors in select types of cancer. Stefanovic et al.[48] talk about how biomarkers change between primary and metastatic tumors, and how the accurate assessment of biomarker conversion between primary versus metastatic can minimize overtreatment for metastatic tumors. In [49] Bhullar et al show that some biomarkers are highly concordant between colorectal cancer and metastatic colorectal cancer, therefore a molecular examination of either a primary tumor or its corresponding metastatic site is enough for designing the individual treatment. Gomez-Roca and his colleagues [50] examined 49 patients to see if the expression of a group of biomarkers - epidermal growth factor receptor (EGFR), vascular endothelial growth factor receptor, Ki-67, and excision repair cross-complementing (ERCC1) - were concordant in non-small cell lung cancer and its metastatic site. They demonstrated that the expression of the evaluated biomarkers is discordant between a primary tumor and its

corresponding metastatic site in 33 percent of cases. In only 18 percent of the tested population, the expression of the biomarkers is the same in both primary and metastatic. Ansari et al. [51]examined 17 cases of primary PDAC and their lymph node metastases to study the expression of Mucin 4 (MUC4) antibody, which is a proposed role in pancreatic cancer progression during the cancer metastasis by comparing its expression in primary versus metastatic pancreatic ductal adenocarcinoma. They noticed that MUC4 was expressed in both primary and secondary tumor with concordance of 82 %.

Therefore, based on the primary tumor, one cannot provide a suitable treatment plan, and based on what we see in the literature, much more research are necessary to evaluate the relationship between primary and metastatic biomarkers; their expression pattern, the amount of expression and the presence or absence of any antibody in both primary and metastatic tumor.

Chapter 3
Image Registration

In image processing, Image registration is the task of aligning two images that are taken from different viewpoints, different modalities or different time instances, involving the transform of one of the two images such that at the end they will be in one coordinate system. For this purpose, one image is used as the reference and the other one as the source image. A geometrical transformation which is a mathematical mapping from points in one image to the corresponding points in another image is calculated based on these two images to be applied to the source image so that these two images align with each other. In pathology, extracting information from different images of the same slide is done by pathologist looking at them one at the time and this is significantly time consuming. The assessment of expression of multiple biomarkers in a single view is possible only when the consecutive images of a slide are aligned [52].

In this chapter, we will first register brightfield thumbnail whole slide images together using the location of fiducial markers which are artificially added landmarks to a slide. For this purpose, first, we need to detect the hashmarks in the images and later register the two images together. Later, we will register immunofluorescence whole slide images using corresponding features in both the reference image and the source image. Feature-based registration techniques extract naturally occurring features from the images rather than relying on the artificial landmarks.

Methods of image registration could be divided based on different criteria. One of the criteria is the nature of transformation [53]. Based on the Geometrical transformation the image registration is divided to 4 types as follows. In rigid image registration only rotations and translations are used in the transformation.

1. Rigid
2. Affine
3. Projective
4. Curved

3.1. Registration using Fiducial Markers

Since a glass slide is manually located inside the tray of a whole slide image scanner, each scan might result in an image with a slightly different [0 0] coordination, consequently coordination of the points will change and this problem will be resolved by registering the consecutive images of a slide together.

A landmark is a recognizable feature that can be found in an image and can be used to match two images in the thumbnail images, fiducial markers are suitable choices to be used in image registration. For this purpose, first these markers should be detected and segmented in both images automatically. Figure 5 shows a sample of the thumbnail immunohistochemistry image that we use for detecting hash marks. As are seen the hash marks here are in the shape of plus signs.

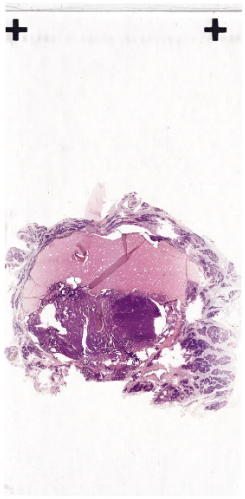

Figure 5. Thumbnail immunohistochemistry image sample

The first method to detect fiducial markers starts with dividing the true color (RGB) thumbnail image to two halves and only the top half is kept for object detection purposes, because in this image the only thing that is needed for object detection and segmentation is the hash mark(s) signs. Then the top half RGB image is converted to grayscale and then to binary simply using Otsu thresholding method (Figure 5). As it can be observed, the white objects in the background are easily discernable, so these objects are labeled and then the properties of objects such as area, perimeter and etc. are calculated, then the boundaries for target objects(blobs) are plotted. This also gives the coordinates of the hashmarks. Finally, the detected hashmarks are imaged. Fig 6-9 and 10-11 present the results of the mentioned steps for finding hashmarks inside the original and rotated images respectively.

Figure 6. Binary image of the thumbnail image

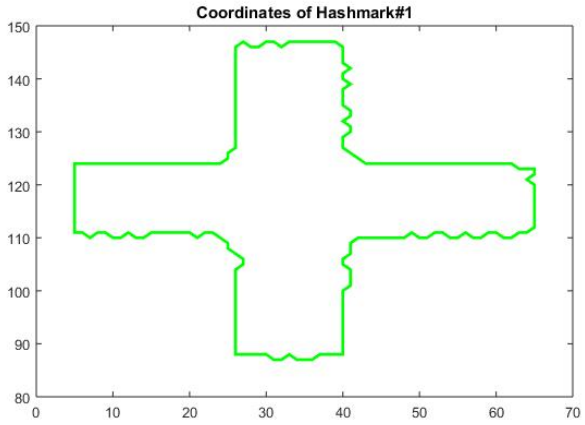

Figure 7. Coordinates of Hash mark number 1

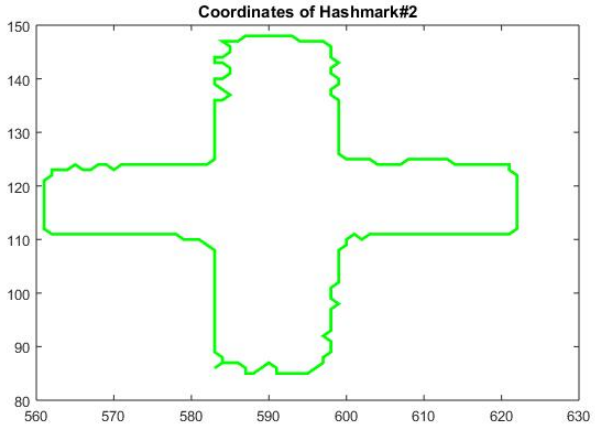

Figure 8. Coordinates of Hashmark number 2

Figure 9. One of the hash marks as an image

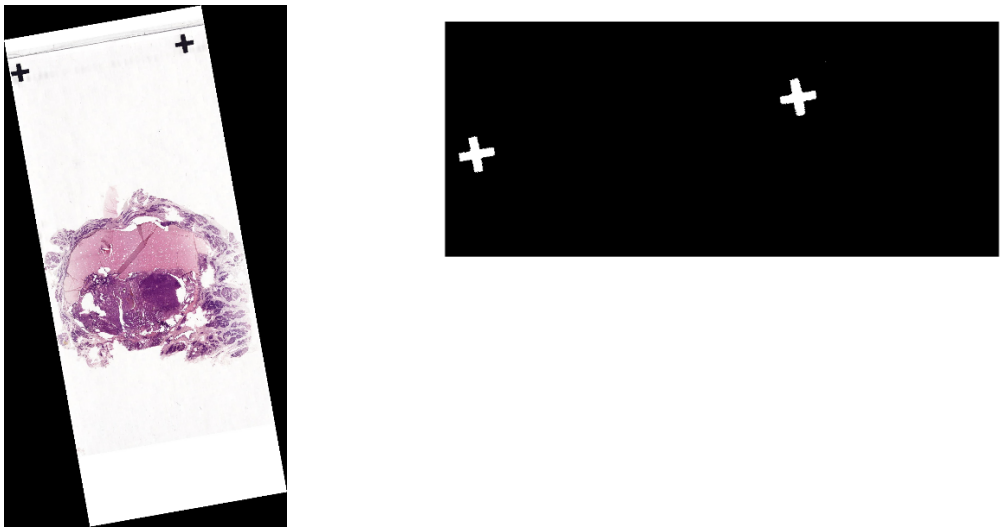

Figure 10. Rotated image(left), Resulted cropped binary image(right)

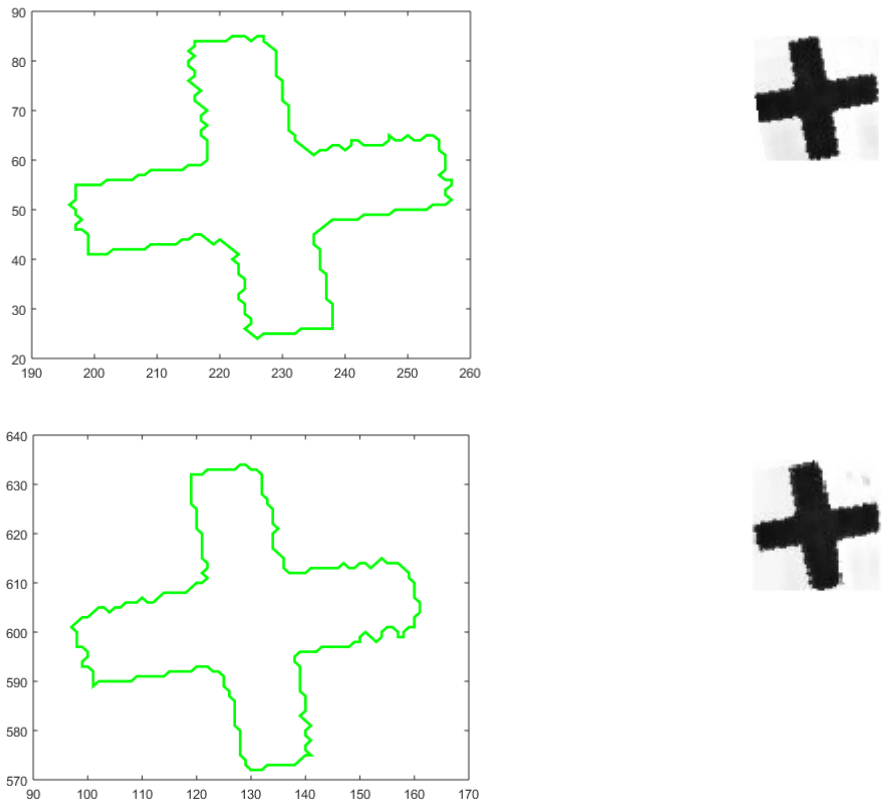

Figure 11. First and second hash mark boundaries and images

The second method for finding hash marks inside of the rotated image is template matching. Template matching is a measure of similarity between a predefined template and a reference image. Templates are usually used to identify small simple objects in a bigger image. There are several methods for template matching such as Naive template matching, image correlation matching and sum of absolute differences [54]. In image correlation matching, the position of the given pattern is determined by a pixel wise comparison of the image with the template that contains the desired pattern. To calculate this comparison, normalized cross correlation is a reasonable choice. Normalized cross correlation (NCC) is calculated between the template and the original thumbnail image [55], the amount is

between -1 and 1 which 1 means the two images are identical and -1 means one image negates the other image and finally 0 means the two images are not correlated at all. In this method the maximum correlation coefficient is at the starting point of the template. Figure 12 shows the rotated image and the detected hashmark.

Figure 12. Rotated image(left) and the template (right)

After the hashmarks are detected, then the coordinates of their corresponding corners are used to calculate the transform matrix and then to align the rotated image(source) to the original one(target). Figures 13-15 show the original thumbnail

image, the rotated thumbnail image and the rotated image registered to the original image respectively.

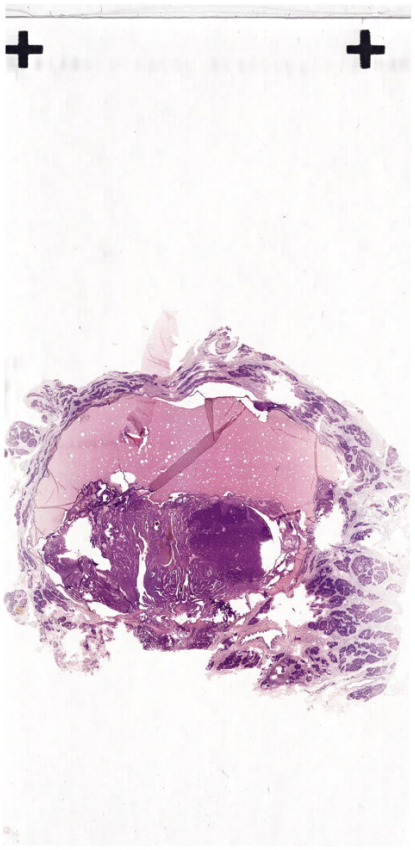

Fig13. Original thumbnail image

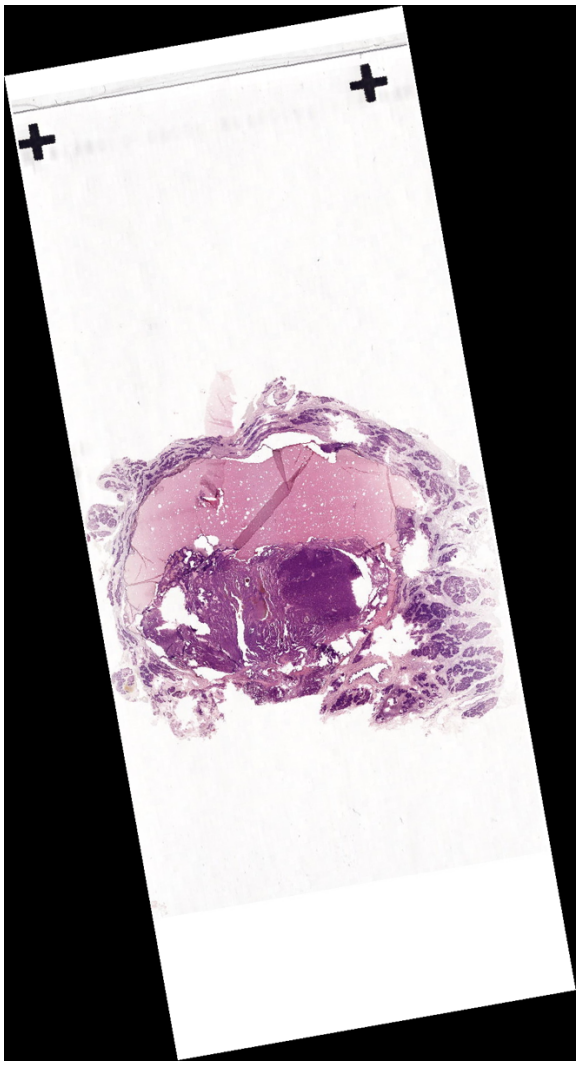

Fig14. Rotated thumbnail image

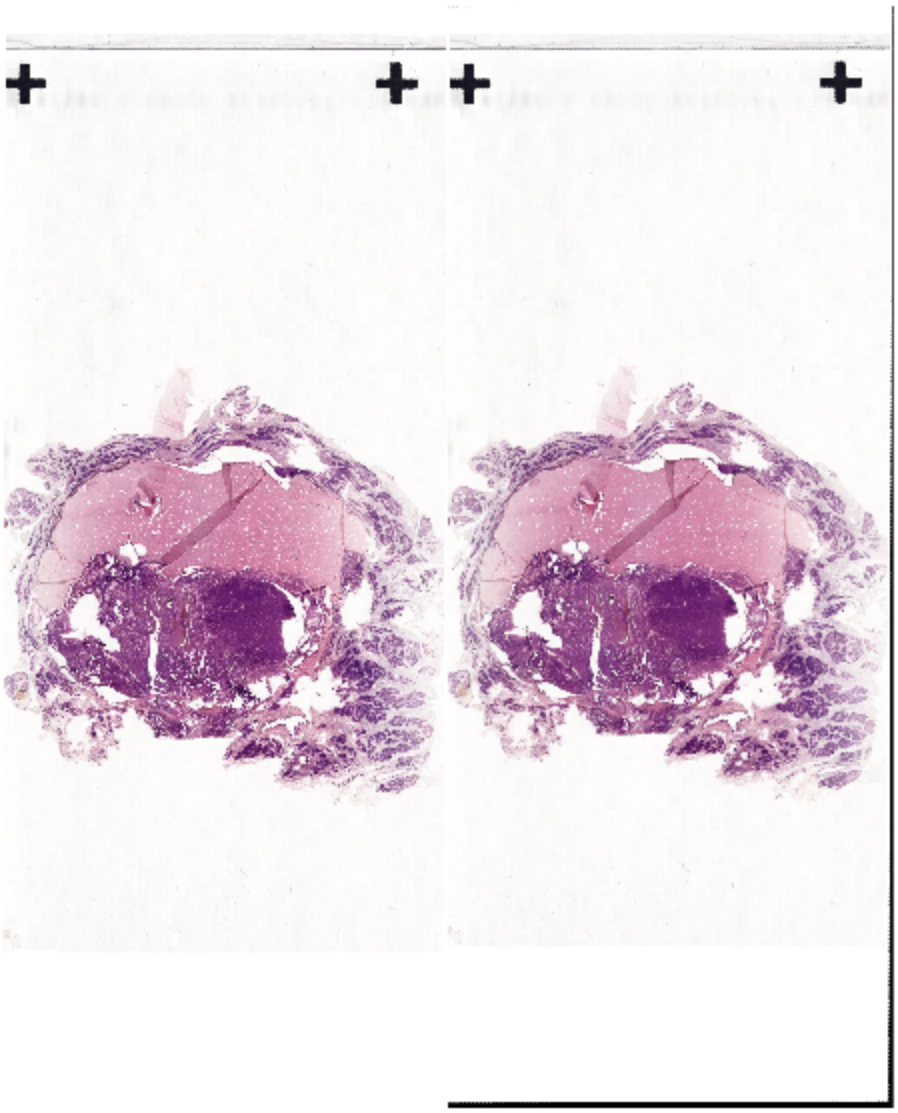

Fig15.Rotated thumbnail image registered to the original thumbnail image

3.2. Image registration in WSI immunofluorescence images

We previously described a method for registration of two thumbnail whole slide images using the coordination of the artificial landmarks. However, a thumbnail image is a low-resolution image that gives only an overview of the high-resolution image. What we have, is a set of immunofluorescence images that need to be registered. Therefore, we developed automatic feature-based registration techniques which extract naturally occurring features from the images rather than rely on artificially created features.

Analysis of the sequentially scanned whole slide images showed that the differences between slides did not include non-rigid deformation from one slide to the other. The differences between sequential images involved changes in intensity, rigid translation, and possibly rotation. Therefore, we developed schemes based on intensity, and on features extracted from the images. The result of intensity-based registration is shown in Figure16.

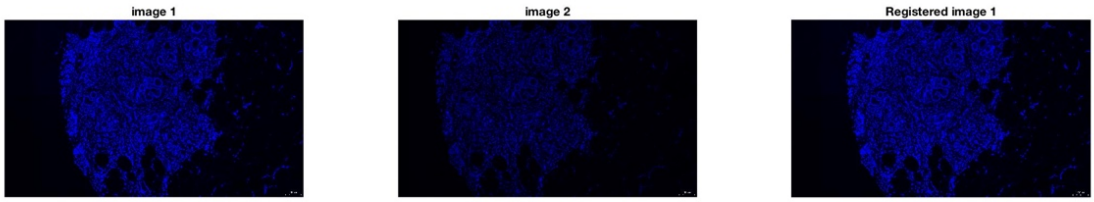

Figure 16. Intensity based registration result

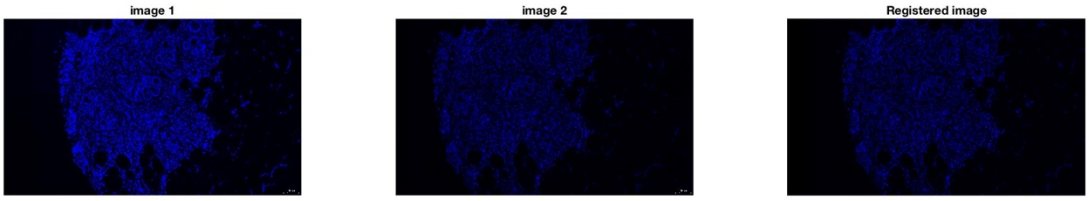

Figure 17. Feature based registration result

To be able to implement feature-based registration, first the features of interest (lines, points, ...) need to be extracted. Even though there are no perfectly circular shapes in the whole slide images, by applying the Hough Transform and limiting the radius of the circles we can identify roughly circular regions of the images which can be used to generate features for the registration process. An example is shown in Figure 18 where the circular regions identified are shown as green circles. The centers of these circles were then used as points of interest for feature-based registration.

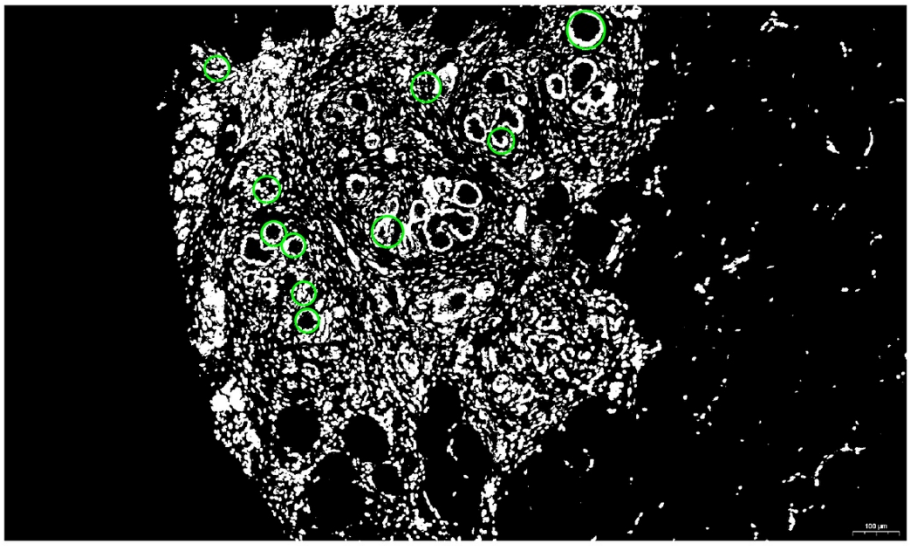

Fig.18 Hough Transform result

The second method used for extracting key points to be used for feature-based registration made use of the regionprops algorithm which finds the centroids and areas of contiguous blobs in an image. The whole side image was binarized as shown in Figure 19.

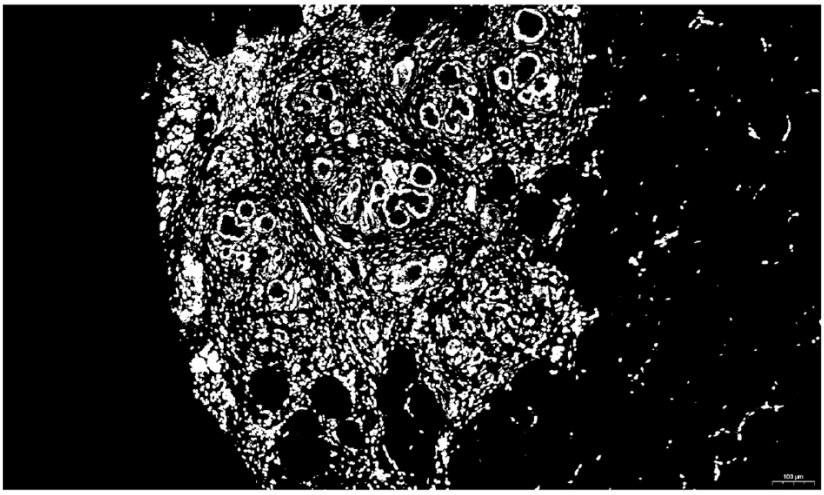

Figure 19. Binarized Image

The binarized image was processed using morphological operations of dilation and erosion. The result is shown in Figure 20. The process results in the filling of "holes" in the image and filtering out some of the noise, resulting in isolated connected regions.

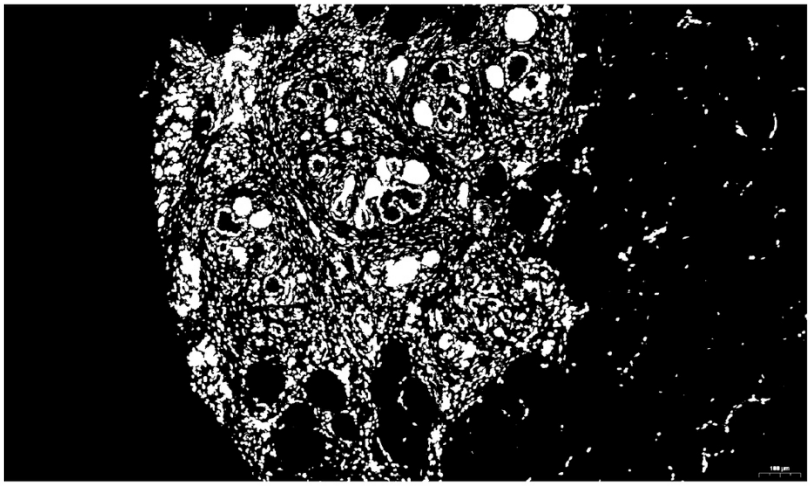

Figure 20. Binarized image after morphological processing

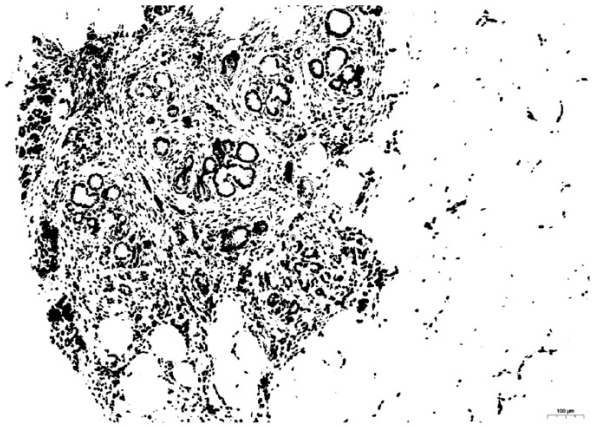

Fig 21. Inverted Image of the processed Binary Image

If we now take an inverse of this processed image as shown in Figure 21, and then take the intersection of the inverted image and the original binarized image we obtain the image shown in Figure 22 that contains only the connected components (blobs) generated by the processing operation.

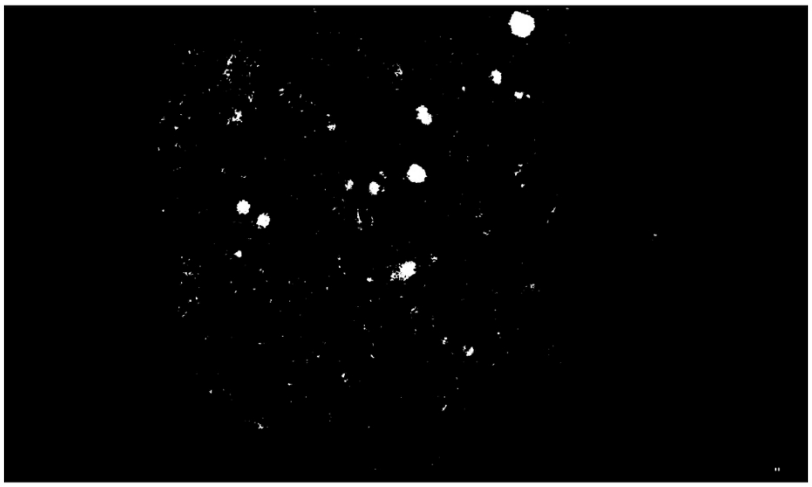

Figure 22. Intersection of the inverse of the morphologically processed image from Figure 21 and the original binarized image from Figure 19

The centroids and areas of the connected components are found, and the three connected components with the largest areas are selected. The centroids of these areas, shown by crosses in Figure 23 are used as features for constructing the affine matrix which is used for registering the images.

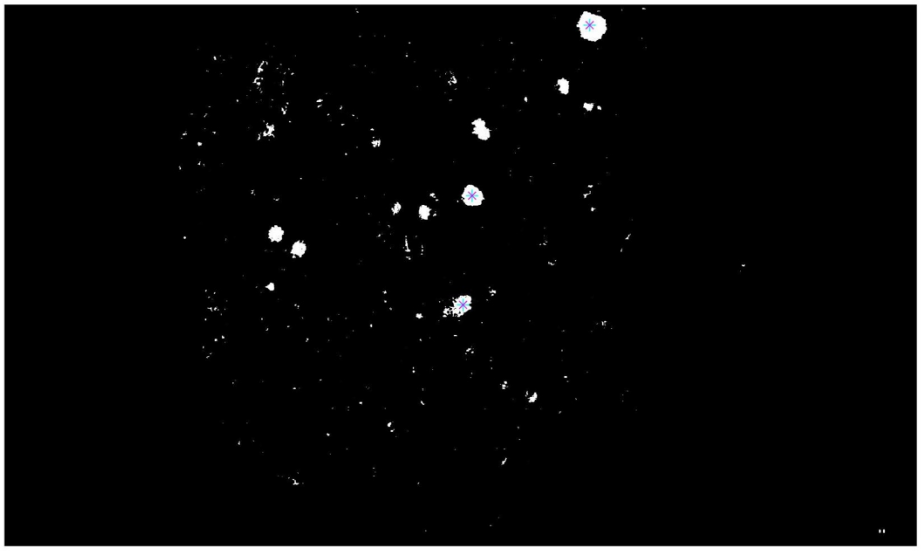

Figure 23. Extracted points for Feature based Image registration are marked

After the registration was completed, color thresholding was used to separate out the regions in the image that had taken up the different antibodies. Recall that the antibodies fluoresce at different colors. An example of an image with two different antibodies is shown in Figure 24. We can view each colored region as a cell or a set of cells that have taken up a particular antibody.

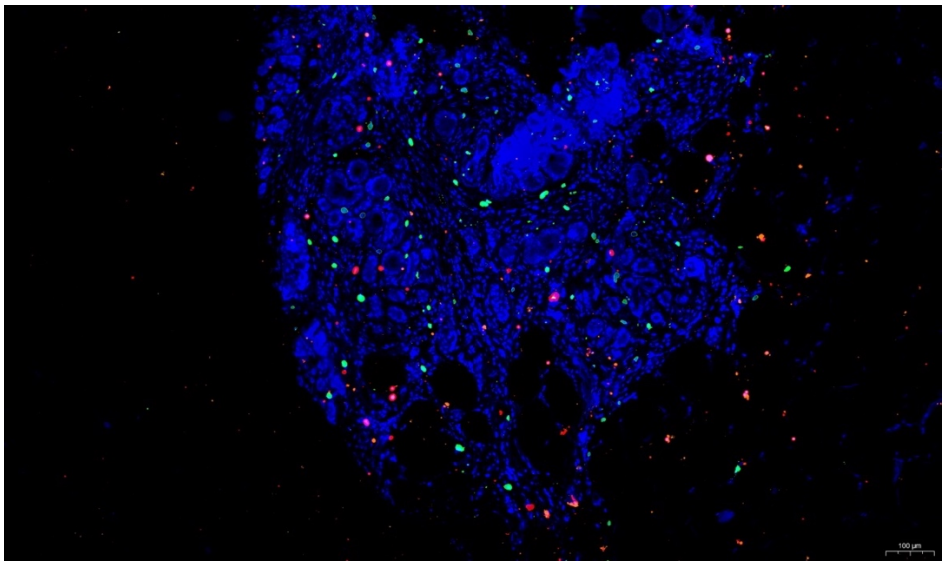

Figure 24. Example of an image with fluorescent labels green and red for two different antibodies.

The following image shows the results of the color thresholding algorithm to obtain the location of the cells which have taken up the antibody labeled with the green fluorescent label.

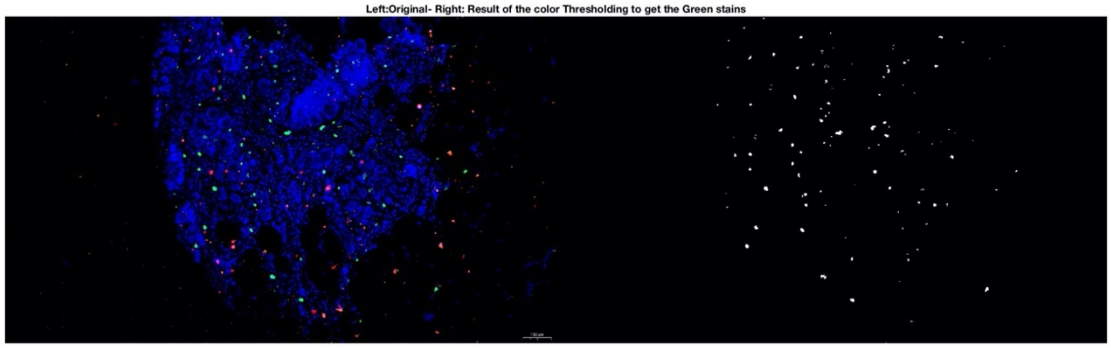

Figure 25. Color thresholding result for the antibody labeled with a green fluorescent label

Each location which takes up a particular antibody is treated as an object and statistics for each object are collected for further analysis. The statistics include the centroid of the objects, their major axis length, minor axis length, the area of every object, which is simply number of pixels for that object, and the perimeter of the object. A sample of this data is presented in Table 1. Note that we also know the coordinates of each of these objects.

Area	Centroid(x)	Centroid(y)	Major Axis Length	Minor Axis Length	Perimeter
25	563.44	308.92	6.908159097	4.773259323	15.341
9	593	617	3.464101615	3.464101615	7.476
15	597.6	664.6	4.760952286	4.188874153	10.751
16	599.5	178	5.977883203	3.596050484	11.884
18	601.666667	84.8333333	5.777885617	4.210362088	12.62
12	619.5	422	4.618802154	3.464101615	9.436
59	624.79661	484.254237	11.82523465	7.344929342	30.318
21	655.571429	40.7142857	6.595402825	4.261516809	13.935
147	665.959184	800.843537	14.75115672	13.07430351	43.253
81	676.753086	144.037037	16.18902405	7.008253874	36.769
99	677.121212	480.575758	12.34046587	10.38219809	33.147
65	676.784615	533.076923	10.16041851	8.579686703	27.151
55	678.054545	307.745455	9.641168609	7.530137069	24.314
18	692	423.5	8	3.464101615	15.622
131	696.992366	735.496183	15.35785436	11.50872129	42.955
9	696	415	3.464101615	3.464101615	7.476
9	697	426	3.464101615	3.464101615	7.476
26	723.615385	7.19230769	7.889735877	4.393894448	16.54
177	731.768362	600.237288	18.2798396	12.85635299	49.414
23	748.869565	254.695652	8.061365453	3.902710368	15.895
12	751.5	240	4.618802154	3.464101615	9.436
78	760.051282	132.974359	12.78769674	8.146835774	30.583
93	784.021505	396.516129	13.57905079	9.05811407	34.048
85	824.576471	448.505882	12.73504546	8.791204916	32.667
65	824.676923	513.907692	11.78267174	7.232278057	27.589
120	832.766667	280.666667	16.35715314	9.53212038	39.027
55	842.490909	179.509091	10.71159917	7.555746268	27.407
77	846.506494	553.181818	15.70405159	7.41622206	38.737
73	853.945205	73.1643836	10.95886725	8.733915833	28.35
47	853.93617	473.702128	9.935543994	7.061943978	25.356
133	858.984962	745.210526	15.88983429	10.82156048	39.862
30	859	626.5	6.92820323	5.773502692	17.276
83	869.650602	620.313253	10.89124318	10.04298462	31.236
81	872.17284	657.271605	11.52393507	9.203321707	30.285
40	896	168.575	9.725854469	5.584301345	21.709
98	924.479592	363.091837	13.83665372	9.234508886	34.164
31	923.580645	609.612903	9.587529376	5.27800052	21.593
23	924.391304	321.565217	6.766533631	4.502862906	14.671
29	928	120.551724	10.24975469	4.174740493	20.212
159	932.201258	907.138365	21.86372059	12.42707264	66.31
9	928	127	3.464101615	3.464101615	7.476

Table 1. A part of the calculated statistics for the green dots after applying the color thresholding algorithm

Speed Up Robust Features (SURF) is another approach in finding and matching features that are locally distinct points in an image to register WSI immunofluorescence images. The original algorithm used for key point detection is called SIFT (Scale Invariant Feature Transform). SURF is mainly inspired by SIFT but is several times faster than SIFT. One of the most critical advantages of SIFT features is that they are not affected by the scale or the orientation of the image. In SIFT algorithm, first, in order to reduce the noise, the image is blurred using Gaussian Blurring methods. By applying Gaussian Blur, minor details are removed from the image, and only information such as the shape and edges are remained. Then several scales of the original image are generated and later are blurred by Gaussian blur. In the next step Difference of Gaussian (DoG) is calculated such that one blurred image of the original image is subtracted from another less blurred version of the original image. In the next step, the key points are localized; the local maxima and minima are found, and later low contrast key points are removed. In order to find local maxima and minima, every pixel in the image is compared with its neighboring pixels and is selected as a key point if its value is the highest or the lowest among its neighbors. Now that the potential key points haveDesktopDe been specified, a final check is essential to choose the best key points; low contrast key points or those close to the edges are eliminated at this point. Now that the robust key points have been selected, an orientation value to each key point should be assigned so that the key point would be rotation invariant. For this purpose, the gradients in x and y directions and later magnitude and orientation for each pixel are calculated.

In the next step, for each key point, a histogram of the magnitude and orientation for the neighboring pixels of that particular key point is created. The peak of the histogram

would be the orientation of that key point. Finally, in the last step, using the neighboring pixels, their orientations, and magnitude, for each key point, a descriptor which is the representation of that specific point and contains the most important information about that key point, is generated [56]-[57]. In order to register two or more images based on SURF algorithm, first using SURF, features are extracted in both the reference and the source images. Then, the transformation matrix between the two images is calculated, and finally, the source image is registered to the reference image. Figures 26-30 show the original images of Dapi 1 and Dapi 2, matching points of Dapi 1 and Dapi 2, and the registered image, respectively.

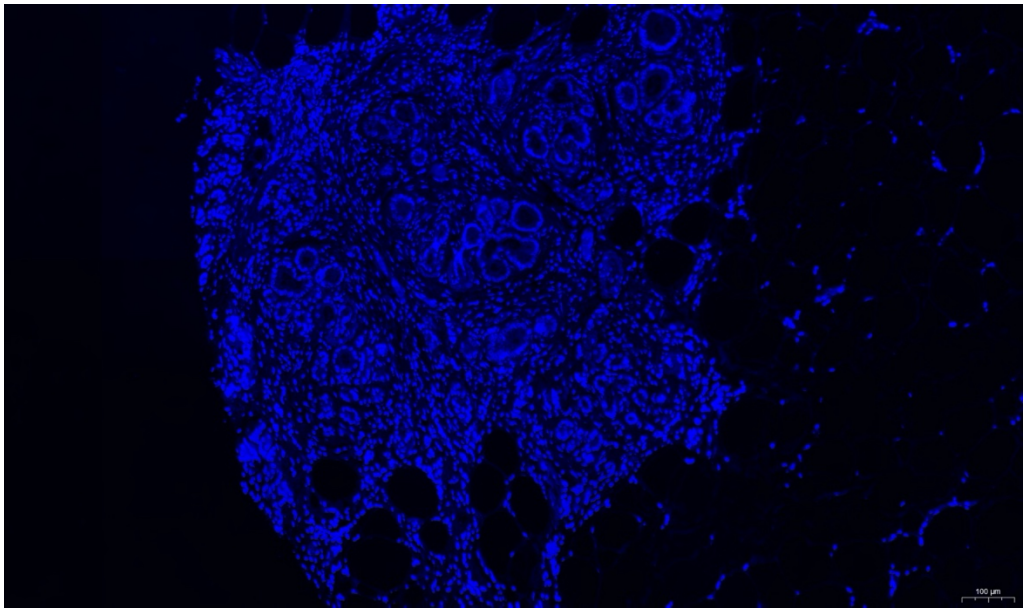

Figure 26. Dapi1

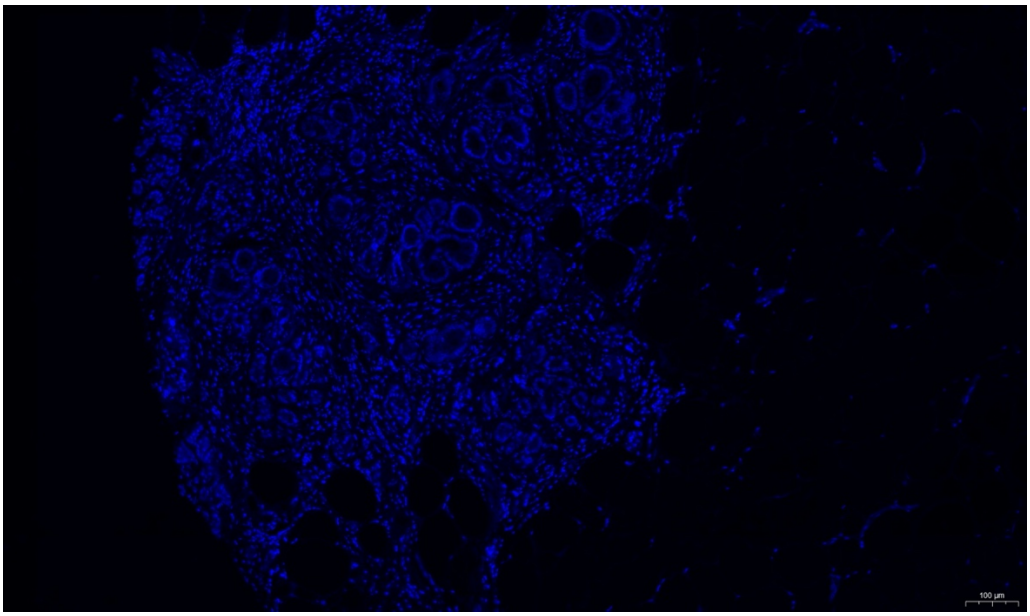

Figure 27. Dapi 2

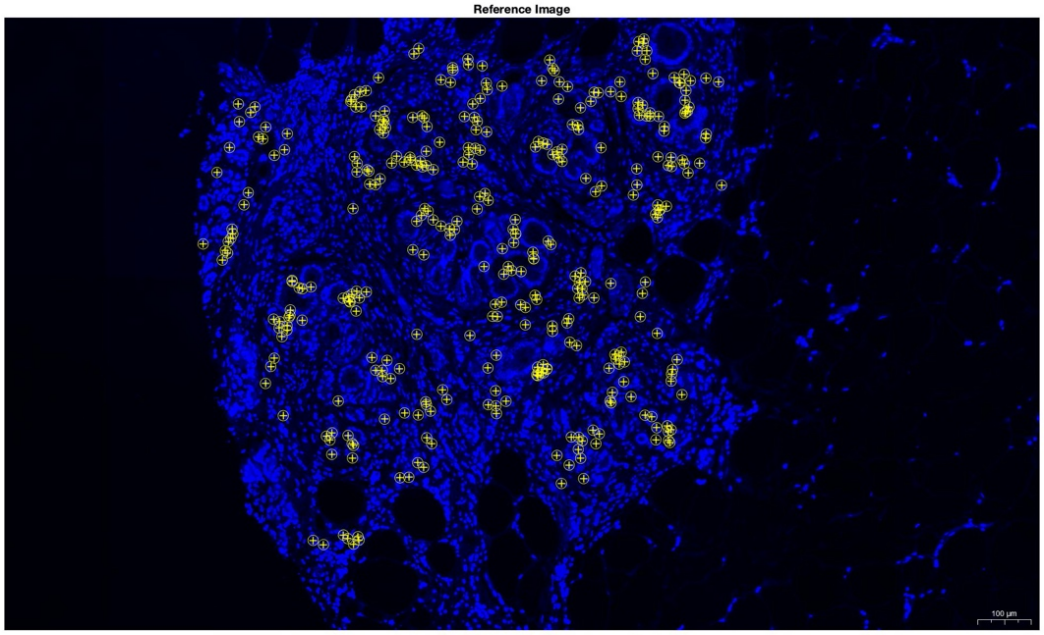

Figure 28. Dapi1 matching points with Dapi 2

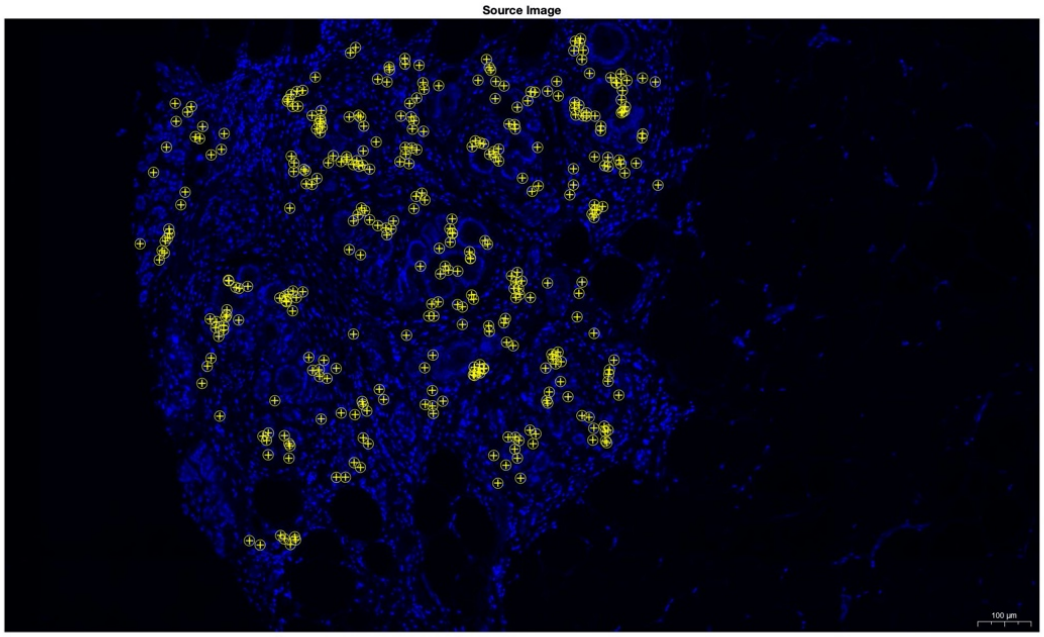

Figure 29. Dapi 2 matching points with Dapi 1

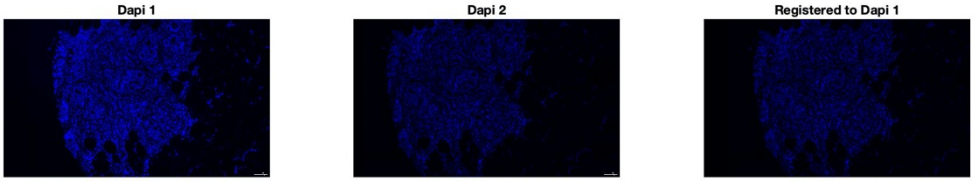

Figure 30. Dapi 2 registered to Dapi 1

In this chapter, we first registered two thumbnail images together based on fiducial markers we detected, using template matching and other methods. Later, we registered immunofluorescence whole slide images together using features in these images. For this purpose, we extracted naturally occurring features in the reference and the source images and then calculated the transformation to register the images together. The registration compensates for moving the tissue slide inside the scanner and brings all the consecutive images of each slide to the same coordinate system. In the next chapter, we will locate multiple biomarkers (antibodies) for each cell and using the antibody vectors, we will design our classifier to investigate the difference between primary and metastatic pancreatic cancer cells.

Chapter 4
Clustering and Classification

Unsupervised learning is a kind of learning that looks for a pattern in the dataset with unlabeled data points. Clustering, which is one method of unsupervised learning, is the grouping of data points into subgroups that share similar properties. This method is a popular technique for statistical data analysis [58].

We deal with millions of pixels in only one sub image of a whole slide image. In order to be able to apply machine learning supervised classification methods on the dataset, unsupervised clustering methods could be an effective way to gain a general overview of the dataset and see how many different classes there are in the dataset.

In this chapter, first using unsupervised clustering methods such as Vector Quantization (VQ) and Principal Component Analysis (PCA), we investigate whether we could cluster the cells into two different groups. This can provide insight about the data we aim to classify ultimately. Later, we design two different classifiers based on morphological features and antibody uptake, using Support Vector Machine (SVM) to examine if we could classify pancreatic cancer cells into primary versus metastatic tumor cells.

4.1. Vector Quantization

To see if we could cluster cells into two categories with different characteristics, we implemented one of the unsupervised clustering methods, LBG Vector Quantization algorithm, on every patient in the dataset.

The preprocessing steps to prepare the images for both clustering methods, including converting the images to binary, are described in the classification methods section in detail. After completing the preprocessing steps, all the images are in binary format, therefore the value of each pixel is either zero or one.

4.1.1. First Method Pixel-Based Vector Quantization

For the first method of pixel-based vector quantization, we started from the first pixel in the first image. Note that the number of primary images is 26 for 26 different antibodies, and the number of metastatic images is 26 for 26 different metastatic antibodies. Each pixel has a coordinate of [x y z], the x and y are length and width coordinates of the pixel in the image and z shows which among the 26 images this pixel is located in. For instance, [150 1103 24] indicates that the pixel is in [150 1103] coordinate in the 24th image. Starting from the first pixel in the first image, its value is either 0 or 1, which demonstrates whether or not that specific antibody exists in the respective pixel. The process continues onto the first pixel in the second image, coding it either 0 or 1, this goes on until the 26rd image. So, for every pixel with an [x y] coordinate, there is now a vector which contains zeros and ones for the absence or presence of antibodies, respectively. An

example could be [1 1 1 0 0 1 1 1 1 0 1 0 0 0 1 1 0 0 1 1 0 0 1]. After preparing all the primary images of a patient for VQ and completing the steps in order to have the primary vectors of that patient, a need for the [x y] coordinates of each pixel was determined. , It was possible at that juncture just to add the coordinates of each vector to the end of that vector and make a matrix vector. Finally, all of these vectors were placed in a long matrix.

At this point, we implemented the Linde-Buzo-Gray (LBG) algorithm, which was introduced by Yoseph Linde, Andrés Buzo and Robert M. Gray in 1980 as a vector quantization algorithm. This method works quite similar to k-means clustering method, the only difference being that the k-means algorithm works with points in each step but in vector quantization it works with accumulated vectors in every step.

For this purpose, LBG vector quantization algorithm was implemented in MATLAB. Additionally, the Python VQ built-in function was used to examine the accuracy of this implemented VQ program. It was observed that both programs' results are close to the same, but the Python VQ built-in function is faster than MATLAB.

All these vectors summed together and would be referred to as a centroid. y0 equals this centroid, and y1 equals centroid plus another random vector. Considering that there are two vectors, y0 and y1. The next steps were making up a matrix, going to each location, taking the distance from y0, finding the ones that are closer to y0, and color-coding them blue for proximity to y0 and red for proximity to y1. Then all points that belong to y0 were averaged. This will be the new y0, then going to each location and taking the distance from y0 and find the ones that are closer to new y0 again, the same process was replicated for y1 too, until no changes were noted anymore [59]-[60].

This VQ process was performed for all primary and metastatic directories. Figure 31 shows that by using an unsupervised clustering method such as vector quantization, no specific pattern that shows any difference between two or more groups of cells was discernible. In the next section, we unpack the second approach to vector quantization that we attempted to see if the cells could be clustered unsupervised.

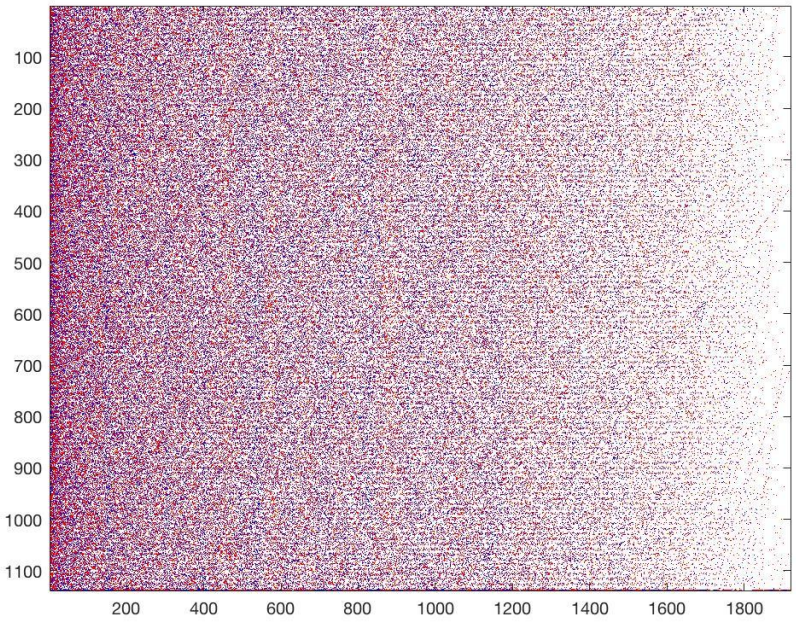

Figure 31. First model Pixel-based clustering using vector quantization

4.1.2. Second Method Pixel-Based Vector Quantization

Preprocessing steps for this part are the same as what had been done in the last part termed first model pixel-based vector quantization. This time, in each image directory, one of the nuclei images is considered as the reference image. The antibody images which are, at that point, binarized, are added together and the summation is saved in the resulting

image named *sum image*. In the reference nuclei image, different nuclei are labeled as different objects. Due to their nature, cancerous cells do not have similar organized shapes, and which has a probable effect on the number of objects and the labeling process.

Starting from the first pixel in sum image, if the pixel value is greater than or equal to one, then an area equal to 50 by 50 pixels around that pixel but in the nuclei reference image is searched in order to find nuclei and calculate the distance from the pixel in the sum image to the found nuclei. The distances from the [x y] coordinates of that pixel to each nucleus in that square are calculated, and the nucleus with the closest distance to that particular pixel is found. In a matrix all non-zero pixels in sum image are associated with a number that is the closest nucleus to that pixel. After this, the vectors associated to each pixel in sum image, will be generated. The length of this vector equals to the number of antibodies in the dataset; each element of this vector is either one or zero which shows the presence or absence of each antibody in that specific pixel respectively. Later all these vectors are concatenated in a long matrix and then all zero vectors are eliminated. After clustering the pixels using vector quantization, all pixels in the same cluster are color coding with the same color and different clusters are color coding with different colors. Figure 32 shows the result for this quantization method. As it can be observed from the result there is no discernable pattern that could distinguish between 2 clusters of cells.

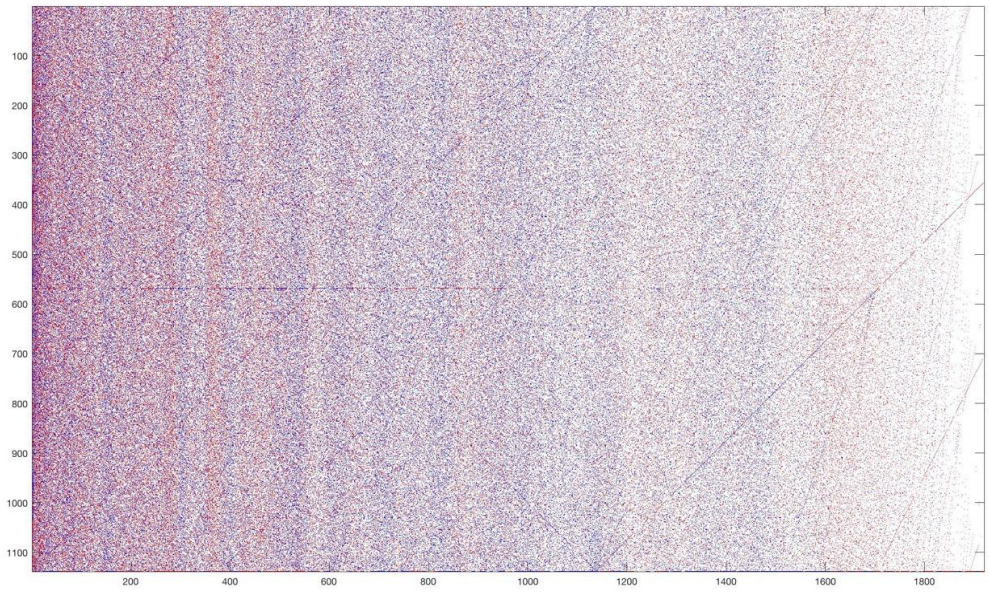

Figure 32. Second model Pixel-based clustering using vector quantization

4.1.3. Object-Based Vector Quantization

After the preprocessing steps, this time instead of investigating pixels for the absence or presence of antibodies, we studied the cells. To see which antibodies were associated with each cell, we found the center of each cell, then defined a vicinity of 50 pixels around the center and checked this circle around the center of every cell to see which antibodies existed in this circle. We then defined a vector for each cell and for each antibody that exists in the cell's neighborhood, with the related element in the vector as 1 and the absence of each antibody as 0. We perform these steps for both primary and metastatic antibodies of a patient, and subsequently perform vector quantization for each one separately. The results are shown in Fig 33 to 40. It can be seen that there is not a specific pattern to distinguish between two or more different classes.

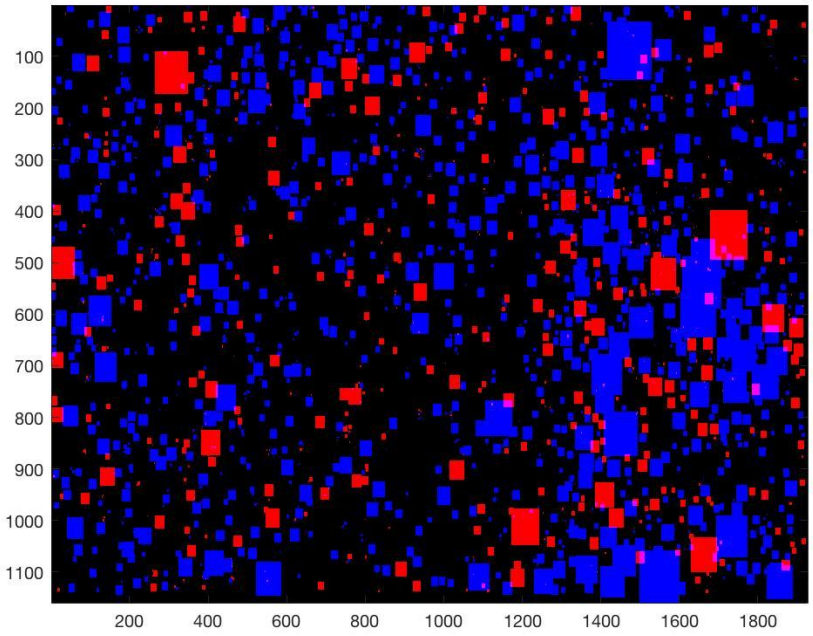

Figure 33. Object-based 2 clusters VQ for Liver Metastasis Patient 80

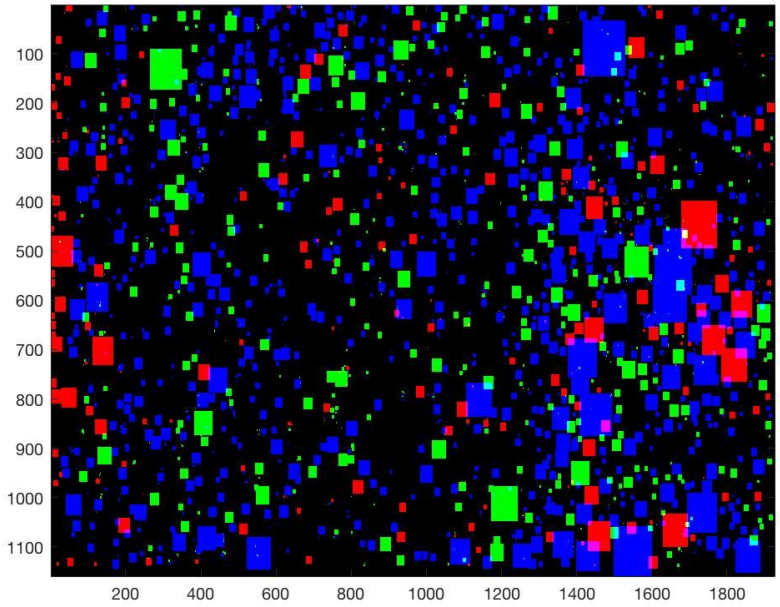

Figure 34. Object-based 3 clusters VQ for Liver Metastasis Patient 80

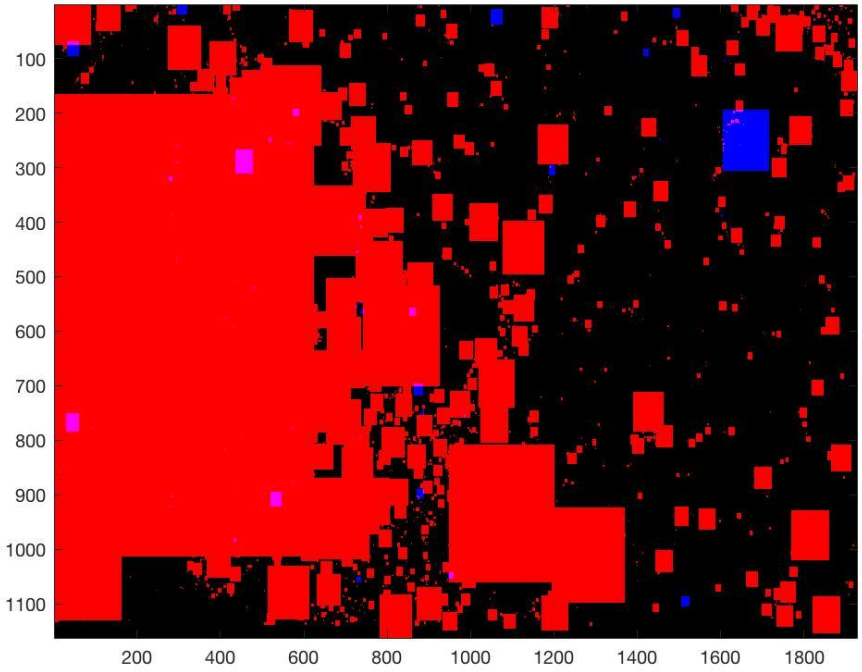

Figure 35. Object-based 2 clusters VQ for Liver Metastasis Patient 105

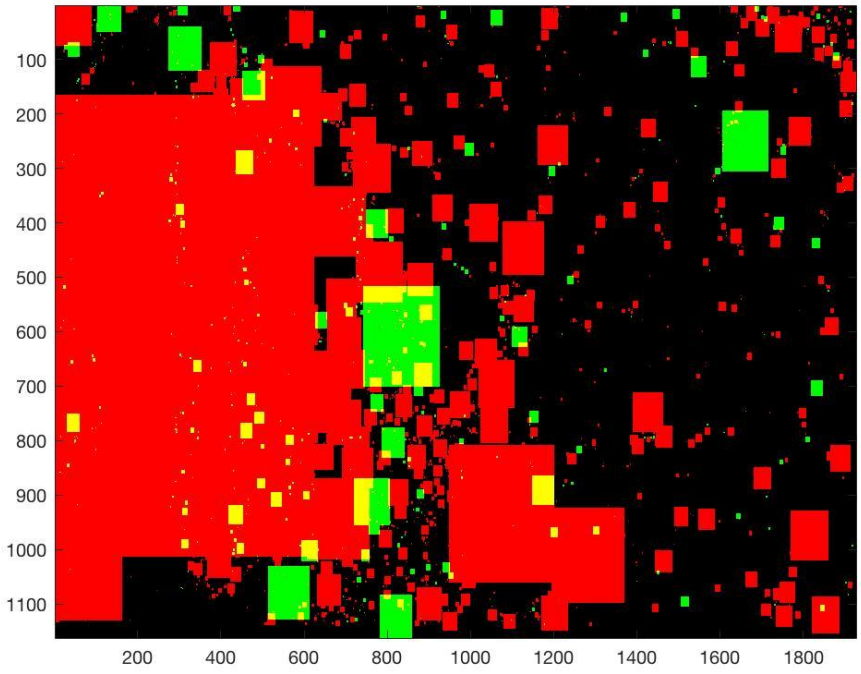

Figure 36. Object-based 3 clusters VQ for Liver Metastasis Patient 105

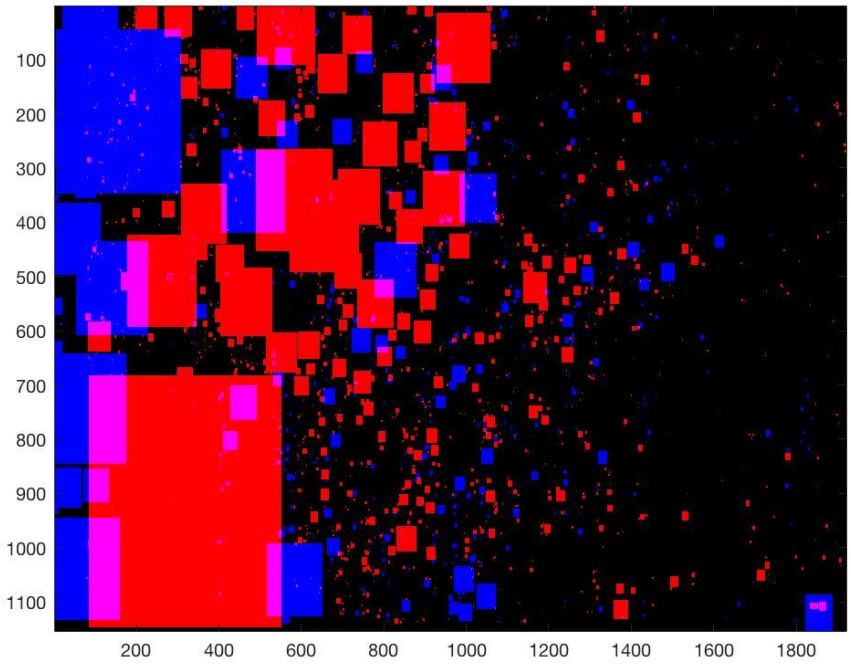

Figure 37. Object-based 2 clusters VQ for Liver Metastasis Patient 8

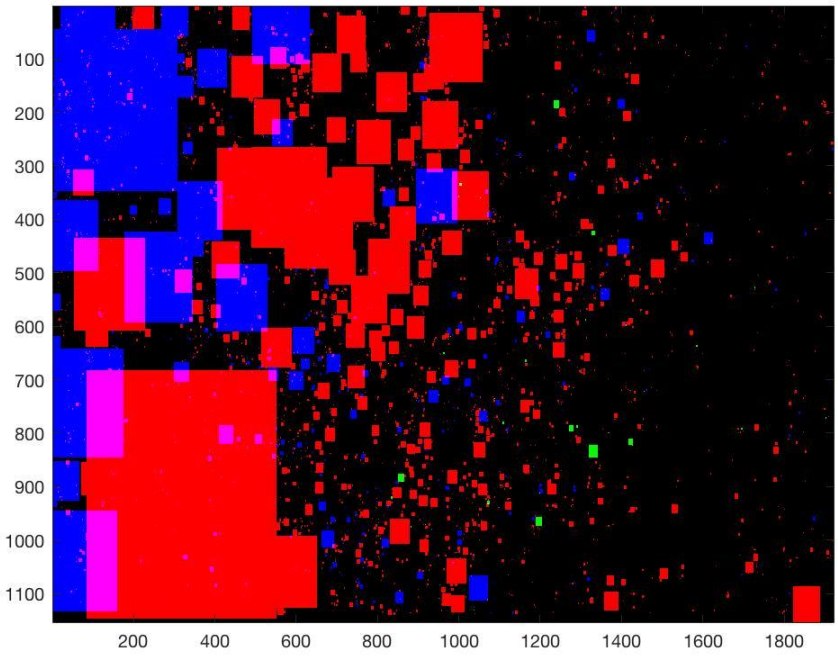

Figure 38. Object-based 3 clusters VQ for Liver Metastasis Patient 8

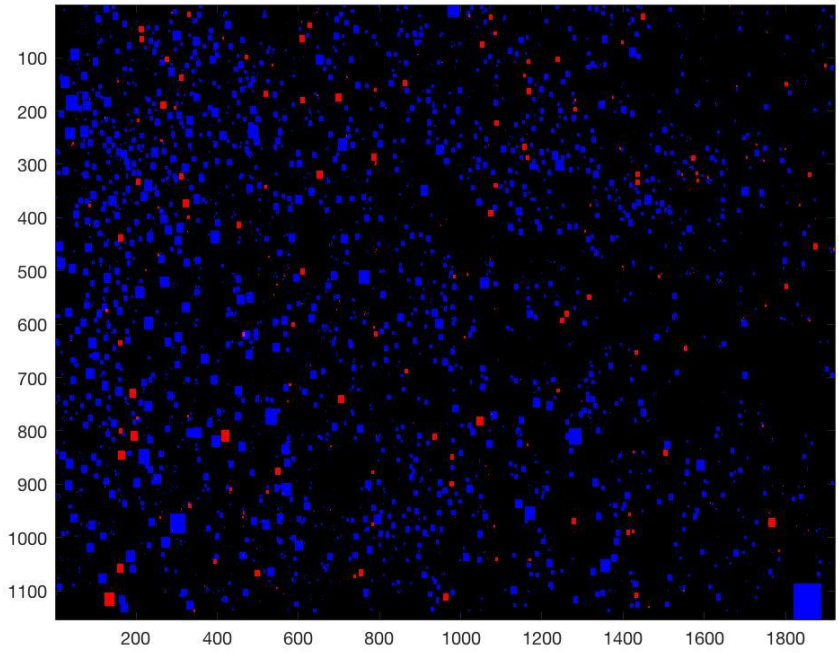

Figure 39. Object-based 2 clusters VQ for Liver Metastasis Patient 57

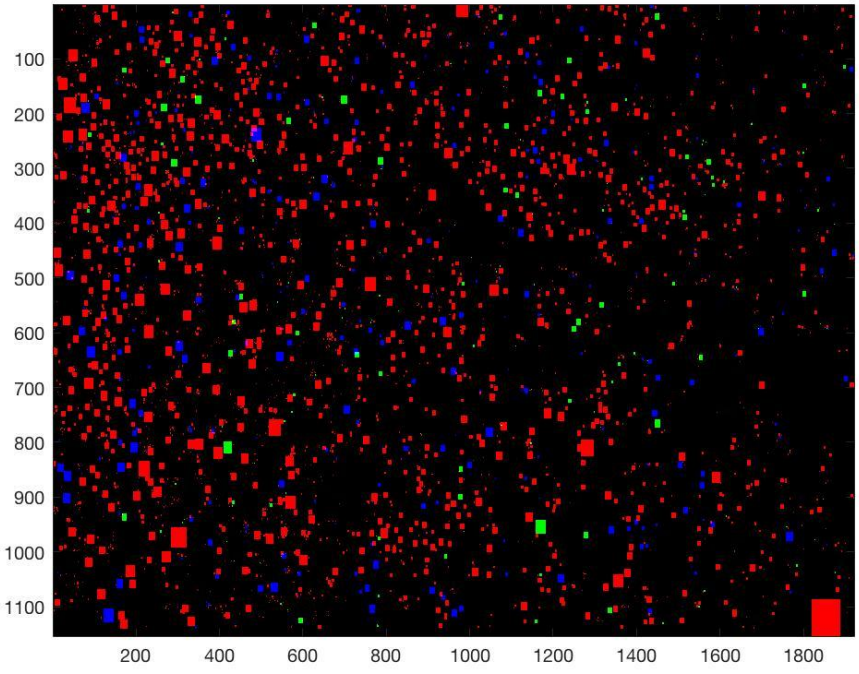

Figure 40. Object-based 3 clusters VQ for Liver Metastasis Patient 57

4.2. PCA (Principal Component Analysis)

PCA or Principal Component Analysis is a technique to reduce the dimensionality of high-dimensional data sets. In fact, PCA projects input data onto a lower-dimensional subspace that still contains most of the information in the large set. It reduces the number of columns while preserving as much information as it can. In PCA, the axes are ranked in order of importance, with differences along the first principal component axis. The first principal component gives the maximum variance, and each succeeding component accounts for as much of the remaining variability as possible [61].

In this project, each variable belongs to a particular antibody, making it a 26 dimensions dataset. in order to look for any insight into the data, PCA is implemented to make undertaking analysis a reasonable task.

Prior to applying PCA, the data had to be prepared first. The images were converted to binaries, the different objects in the Dapi image were then labeled. The centroid of every object was calculated. A surrounding window around the centroid of each object was defined. Then in each antibody image, 26 in total, the coordinates of each object were located, and the surrounding window was applied and searched for that particular antibody. For each object, a vector was specified in which the presence of each antibody appeared as 1 and the absence of that antibody appeared as zero. Once all the vectors were calculated for primary images of a patient, then all these steps were repeated for the metastatic images of that particular patient. Next, all the primary vectors followed by metastatic vectors were put in a long matrix. Finally, Principal component analysis was implemented in MATLAB. Only for two patients, 80 and 105, was the primary and metastasis data fully available. In

other patient cases, either primary or metastasis data exists in the dataset, but not both. Therefore, in order to attempt PCA for other patients, primary data of one patient concatenated to a metastasis data of another patient. Figures 41 to 43 show the result for a few numbers of patients.

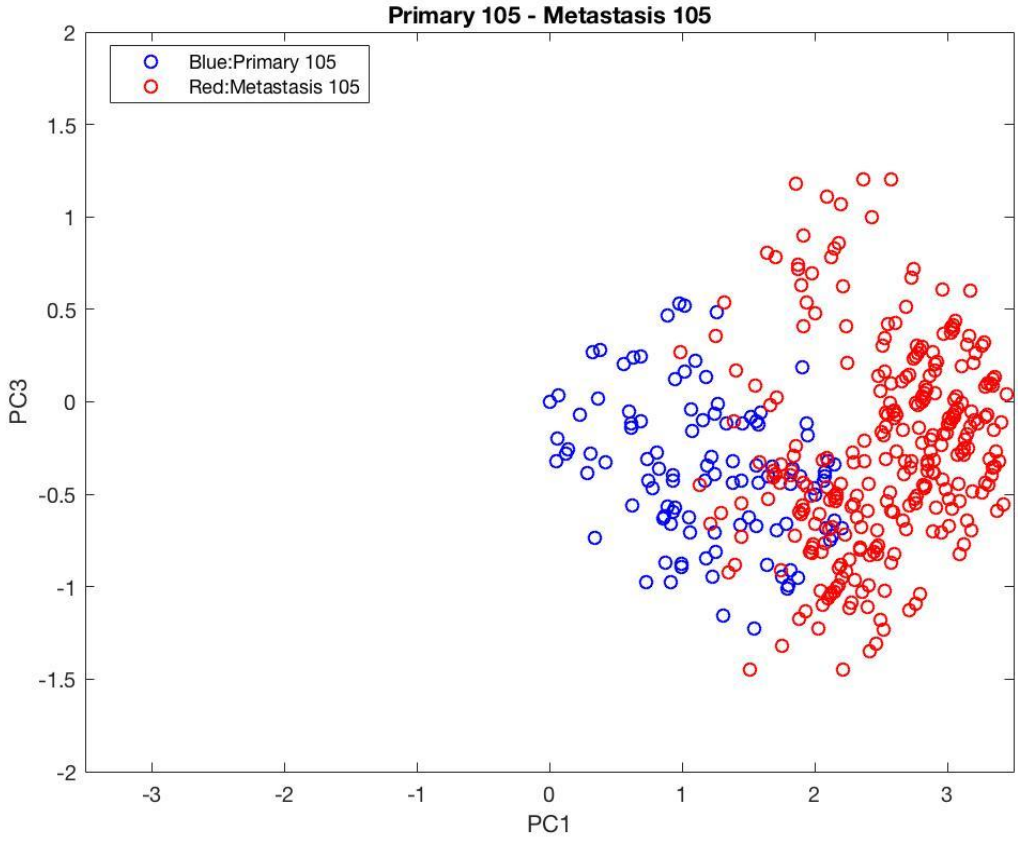

Figure 41. PCA result for Primary Metastasis data of patient 105, PCs 1 and 3

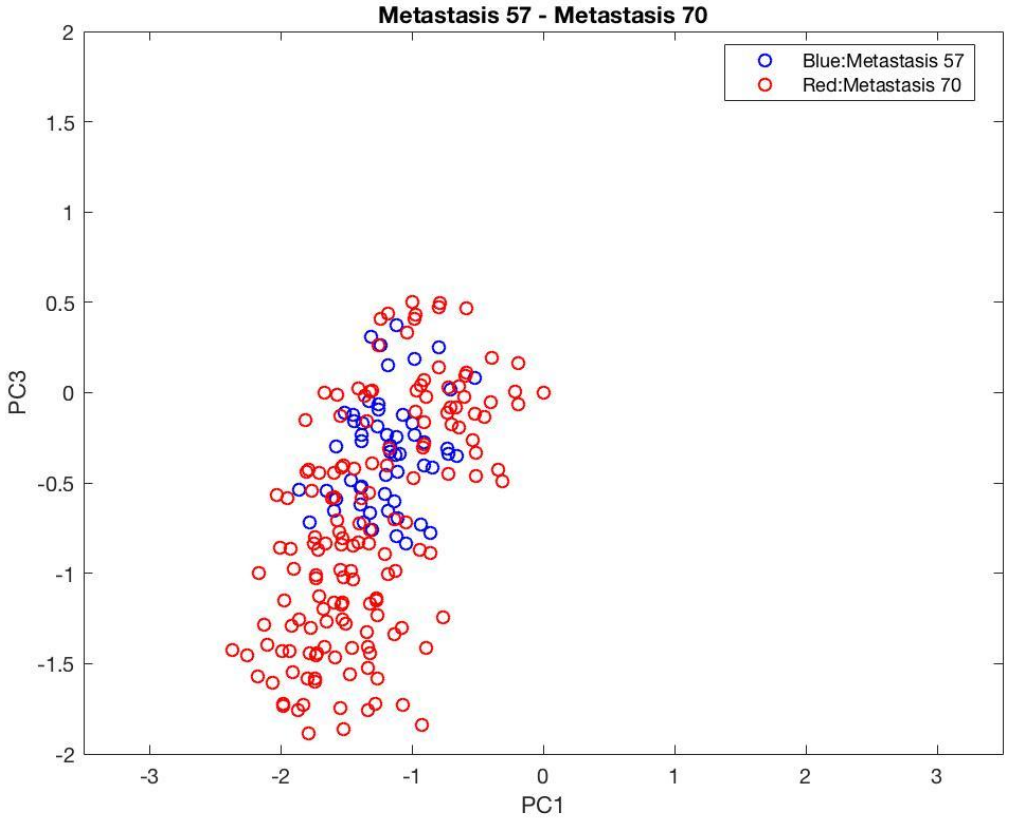

Figure 42. PCA result for Metastasis data of patients 57 and 70, PCs 1 and 3

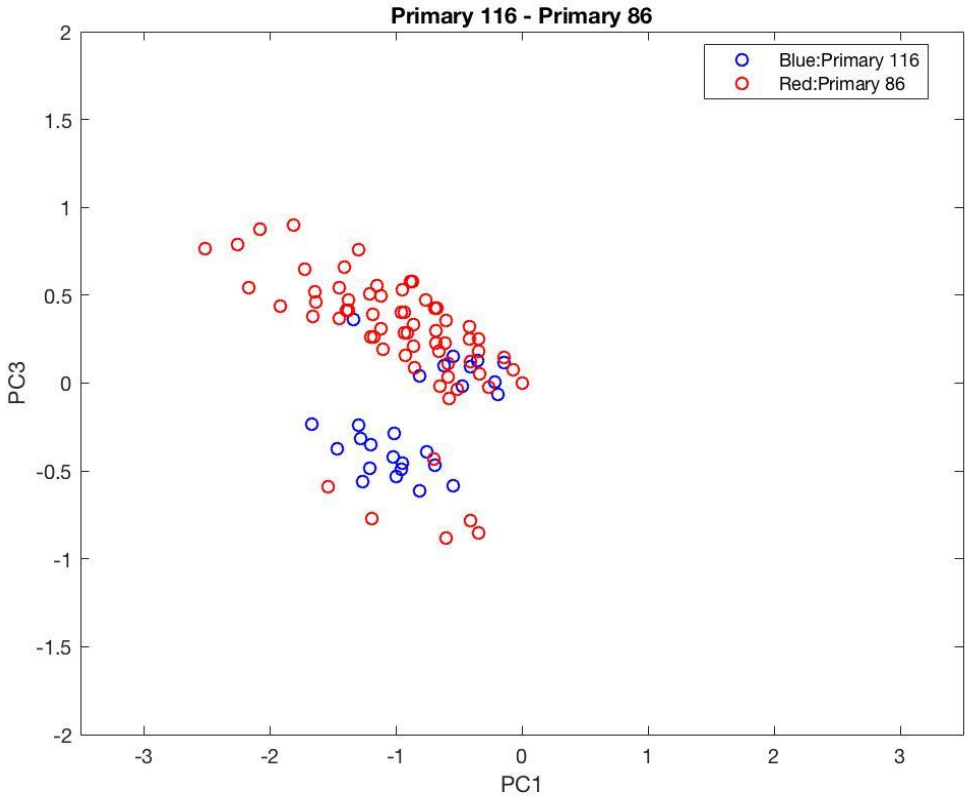

Figure 43. PCA result for Primary data of patients 116 and 86, PCs 1 and 3

As it can be noticed from the results, in cases when the data is both primary or both metastatic but from different patients, PCA could not cluster the data to two different groups. However, when the dataset is primary-metastasis data from one patient, PCA has clustered the data to almost two different clusters. Nevertheless, the results overlap in some cases and this makes them not very reliable. Most of the variance is explained by PCs 1 and 3 -- the first and the third PCs.

4.3. Classification Methods

In a classification problem which is a supervised method, a dataset is divided into two classes for a binary classification or more for a multi-class classification, based on specific features. Features are properties or characteristics of the data that are the same in each class or sub population of a dataset [62]-[63]. Selecting proper features are extremely important as they can reduce the dimensionality of the data, exclude present attributes in the data and also help the classifier to make good predictions. Using a support vector machine, we designed three approaches for classification to see if it were possible to classify cells of a pancreatic cancer tissue into two classes of primary or metastasis using immunofluorescence whole slide images of pancreatic tumors.

4.3.1. Support Vector Machine

Support vector machine (SVM) is a supervised machine learning method that originally debuted in the 1990's in the work of Vapnik and Chervonenkis. First, it was only for linear classifying; however, later in 1992, Vapnik, Boser and Guyon suggested a way to a non-linear classifier [64]. SVM is among the fastest classifiers which can be used in two forms of linear and nonlinear. In the linear form, the classifier is able to divide the data into two categories by a line, while in the non-linear form, the dataset cannot be separated by a line and therefore other kinds of kernels such as Radial Basis Function (RBF) or SIGMOID function are used. Linear SVM uses a hyperplane to divide the data into two classes. This hyperplane is found such that to maximize the distance between the support vectors which are the closest points to the hyperplane. Figure 44 represents the linear SVM problem [65]. For cases in which the dataset is not easy to be separated linearly, the non-

linear SVM works by mapping the dataset to a higher dimensional space where the data is separable. Figure 45 [66].

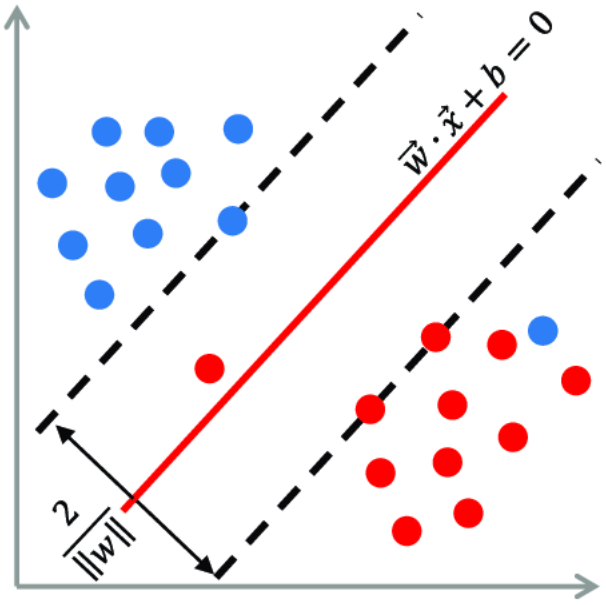

Figure 44. Linear SVM [65]

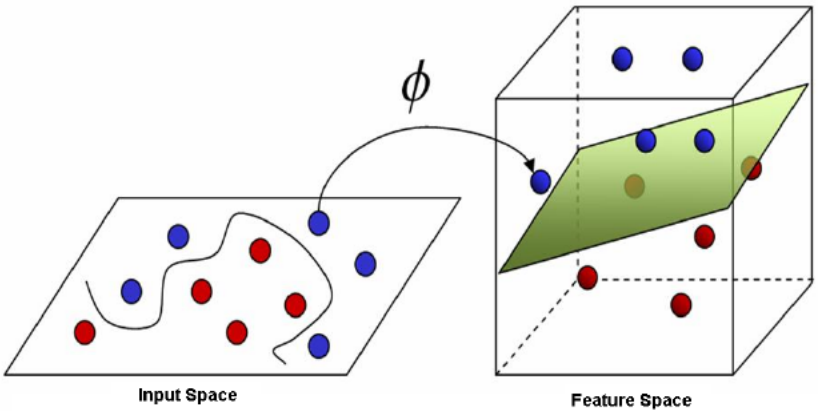

Figure 45. mapping the dataset to a higher dimensional space [66]

4.3.2. Dataset for the classification

Immunofluorescence whole slide images of both primary tumor (pancreas tumor) and secondary tumor (either lung or liver) stained with different antibodies for patient 80 and 105 exist. For other patients, the information of either the primary tumor or the secondary one exists. The dataset is very small, as the public datasets of cancers are mostly brightfield microscopic images and the number of whole slide images are very limited and mostly proprietary. Despite this challenge, we developed the model and tested it on our small dataset.

4.3.3. Preparing the dataset for the classification

In order to prepare the images for two methods of classification in this project, i.e. cell morphology classification and antibody uptake classification sub-method 1, we needed to convert the color RGB images to binary. There are several methods for performing the binarization, however, not all of them give acceptable results. In this project we tried different methods for binarization such as OTSU Global thresholding, which is applied to all pixels in an image and it is based on the histogram of the gray scale version of that particular image [67]. If g(x,y) is the binarized image of f(x,y), the relation between g and f is an equation (1). The other method for binarization is adaptive or local thresholding,

$$g(x,y) = \begin{cases} 1 & \text{if } f(x,y) \geq T \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

in which different thresholds are calculated for smaller parts of an image. With examining several methods for binarization and evaluating the results, we determined that

the best results were obtained by using the watershed segmentation algorithm. The idea of Watershed segmentation which is a region-based approach comes from geography in which a piece of land is divided into several small parts when it is flooded by rain. The watersheds are in fact the lines that divide the land into several smaller parts. The original algorithm of watershed segmentation was proposed by Digabel and Lantu´ejoul and improved by Beucher and Lantu´ejoul [68]. The images are converted to binary using watershed segmentation and the result for one image is presented in figure 47.

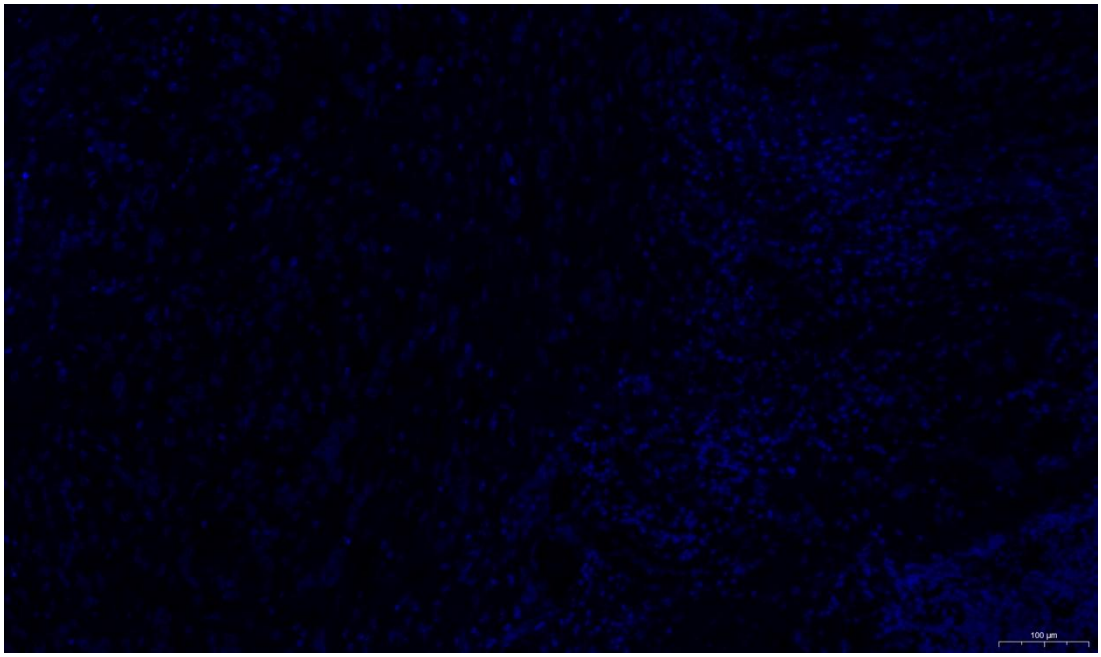

Figure 46. Patient 80 Primary Dapi

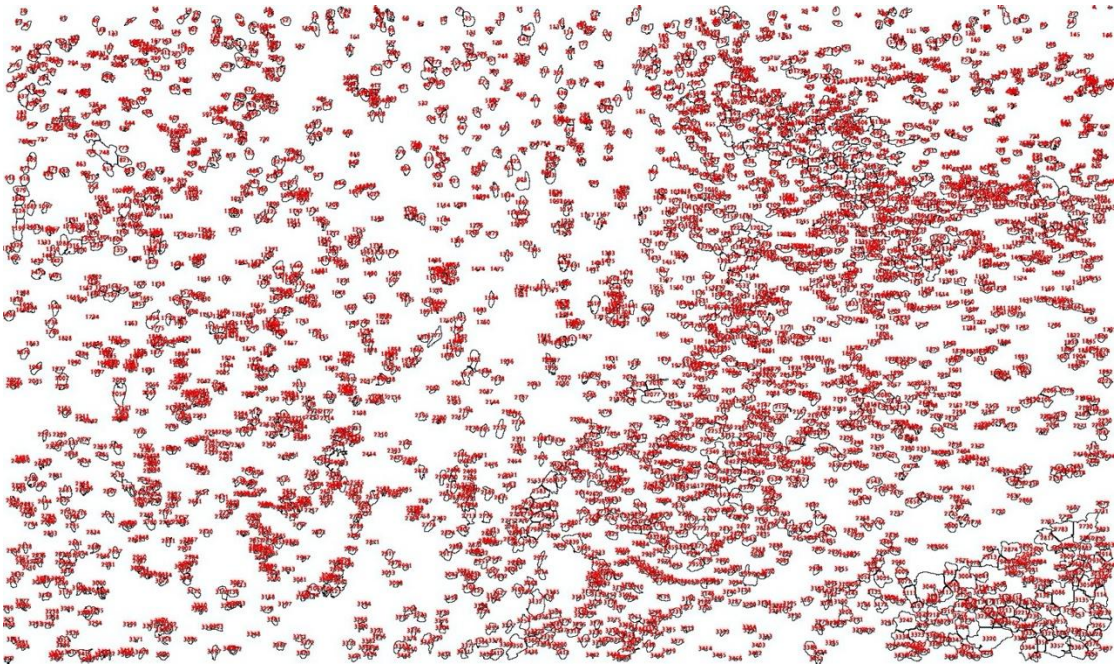

Figure 47. (Watershed Segmentation) for separating the cells (Patient 80 Primary Dapi)

The images are now preprocessed. Altogether, there are 26 antibodies in the different images of primary and metastatic tumors. The list of antibodies are as follows:

Bdac1, ccr3, cd103, cd11b, cd141, cd163, cd19, cd31, cd4, cd54ra, cd56, cd621, cd8, epcam, foxp3, gfap, gzmb, il10, il17, ki67, lox1, mct, mpo, muc1, prg2, sma

For shaping the classification problem and solving that, two types of features are extracted from the immunofluorescence whole slide images of pancreatic cancer of the patients in the dataset. The first type of features is based on the uptake of each cell of different antibodies in the primary and metastatic tumors. The second type of features is based on morphology of the cells including area, perimeter, etc. The three methods of classifying the cells are presented in the following sections.

4.4. Different methods of classification using SVM

4.4.1. Classification based on antibody uptake in binary images

In this method, after converting the RGB images of the cells (dapi) and the antibodies to binary, with each antibody in a different fluorescent image--, we found the center of each cell and defined a vicinity of 50 pixels around the center to check this circle around the center of every cell to see which antibodies exist in this circle. Then we define a vector for each cell. For each antibody that exists in the cell's neighborhood, the related element in the vector was given a 1; while in the absence of each antibody the element in the vector would be given a 0. These steps were performed for both primary and metastatic antibodies of all patients. Therefore, for each cell there will be a vector of 26 elements which these elements would be either zero or one. In the end, there will be a matrix with all primary tumor vectors followed by all metastatic tumor vectors.

The primary cells are labeled as class 1 (C1) and the metastatic tumor cells are labeled as class 2 (C2). This long matrix would be fed to SVM. An example of the data that is fed to this model is presented in the following table.

Primary tumor cells(objects)		**Antibody 1**	**Antibody 2**	**Antibody 3**	**…**				**Antibody26**	**Class**
	Object 1	0	1	1	…	0	0	1	1	C1
	Object 2	1	0	1	…	0	1	1	0	C1
	Object 3	0	0	0	…	0	1	1	0	C1
	...				…					C1
	Object 1350	1	1	1		1	1	0	1	C1
	Object 1351	0	1	0	…	1	0	0	0	C1
Metastatic tumor cells(objects)		**Antibody 1**	**Antibody 2**	**Antibody 3**	**…**				**Antibody26**	
	Object 1	1	1	1	…	1	0	0	0	C2
	Object 2	0	0	0	…	1	1	1	1	C2
	Object 3	0	0	0	…	1	1	1	0	C2
	…									C2
	Object 1795	1	1	0	…	1	1	0	0	C2
	Object 1796	1	1	1	…	0	0	1	1	C2

Table 2. Classification- Antibody uptake in binary images for patient 105

4.4.2. Classification based on antibody uptake in grayscale images

In this method, only the images of the cells (dapi) are binarized. The antibody images are converted to grayscale images and then the intensities are normalized. When searching the vicinity of 50 pixels around each cell for existing antibodies, the normalized

amount of the intensity of the antibody is put in the related element in the antibody vector instead of just a numerical one. So, for each cell, there will be a vector of 26 elements. At the end, there would be a matrix with all vectors of primary tumor antibodies followed by all vectors of metastatic tumor antibodies. The primary cells are labeled as class 1 (C1) and the metastatic tumor cells are labeled as class 2 (C2). This file is fed to SVM. An example of this model is presented in the following table.

Primary tumor cells(objects)		**Antibody1**	**Antibody 2**	**Antibody 3**	**…**			**Antibody26**	**Class**
	Object 1	0.07	0.2	0.02	…	0.05	0.003	0.019	C1
	Object 2	0.0016	0	0	…	0	0.006	0.0003	C1
	Object 3	0.01	0.003	0	…	0	0	0.005	C1
	...				…				C1
	Object 1350	0.045	0.0005	0.0032		0	0	0.033	C1
	Object 1351	0.014	0.012	0	…	0	0	0.055	C1
Metastatic tumor cells(objects)		**Antibody1**	**Antibody 2**	**Antibody 3**	**…**			**Antibody26**	
	Object 1	0.022	0.0014	0.013	…	0	0	0.0067	C2
	Object 2	0.0013	0.044	0.065	…	0	0	0.0018	C2
	Object 3	0.0019	0	0.0061	…	0	0	0.009	C2
	…								C2
	Object 1795	0	0.006	0.018	…	0	0	0.0444	C2
	Object 1796	0.0012	0	0	…	0	0	0.0017	C2

Table 3. Classification- Antibody uptake in grayscale images for patient 105

4.4.3. Classification based on Morphological features

The second type of features is based on morphology of the cells. Morphological image processing are non-linear operations related to the shapes of objects in the images. The purpose of this part is to investigate if based on the morphology of the cells the classifier could differentiate between primary and metastatic tumor cells. An example of this model is presented in the following table.

Primary tumor cells(objects)		**Area**	**Perimeter**	**Solidity**	**…**			**Eccentricity**	**Class**
	Object 1	42	25.16	0.5368	…	0.05	0.003	0.5873	C1
	Object 2	32	25.67	0.8421	…	…	0.06	0.9409	C1
	Object 3	…	…	…	…	…	…	…	C1
	...				…				C1
	Object 1350	…	…	…		…	…	…	C1
	Object 1351	11	8.79	1	…	…	…	0.6847	C1
Metastatic tumor cells(objects)		**Area**	**Perimeter**	**Solidity**	**…**			**Eccentricity**	
	Object 1	20	14.6	0.73	…	…	…	0.86	C2
	Object 2	8	13.72	0.99	…	…	…	1	C2
	Object 3	…	…	…	…	…	…	…	C2
	…								C2
	Object 1795	…	…	…	…	…	…	…	C2
	Object 1796	18	14.2	0.89	…	…	…	1	C2

Table 4. Classification-Morphology features for patient 105

4.5. Results of the Classification

4.5.1. Classification- Antibody uptake in binary images

In this case, the highest accuracy reached was 82%. The classifier was trained with the primary and metastatic information of a patient and then tested with the information of another patient. In this method, the antibody vectors that have been generated for primary and metastatic tumor cells specify the information about presence or absence of the antibodies and inform whether or not they belong to the cell. In fact, metastatic tumor cells are just primary tumor cells that immigrate through the blood or lymph systems to other places in the body. There is no typological difference in between primary tumor cells and metastatic tumor cells [69] only a differing amount of protein expression in some cases.

Therefore, putting one for the presence of the antibody in the vicinity of the cell and putting zero for the absence of the antibody does not clarify if the protein expression of a specific antibody is different for primary versus secondary tumor cells.

In this method, the classifier classified the cells(objects) mostly correctly when the borders of the cells were clearly specified, and the cells were separate. Therefore, this classifier works under the circumstance of clear borders for nuclei. Figure 48 to 50 show the results for metastatic tumor of patients 86, 105 and 8, respectively. The blue objects are the cells that have been classified correctly.

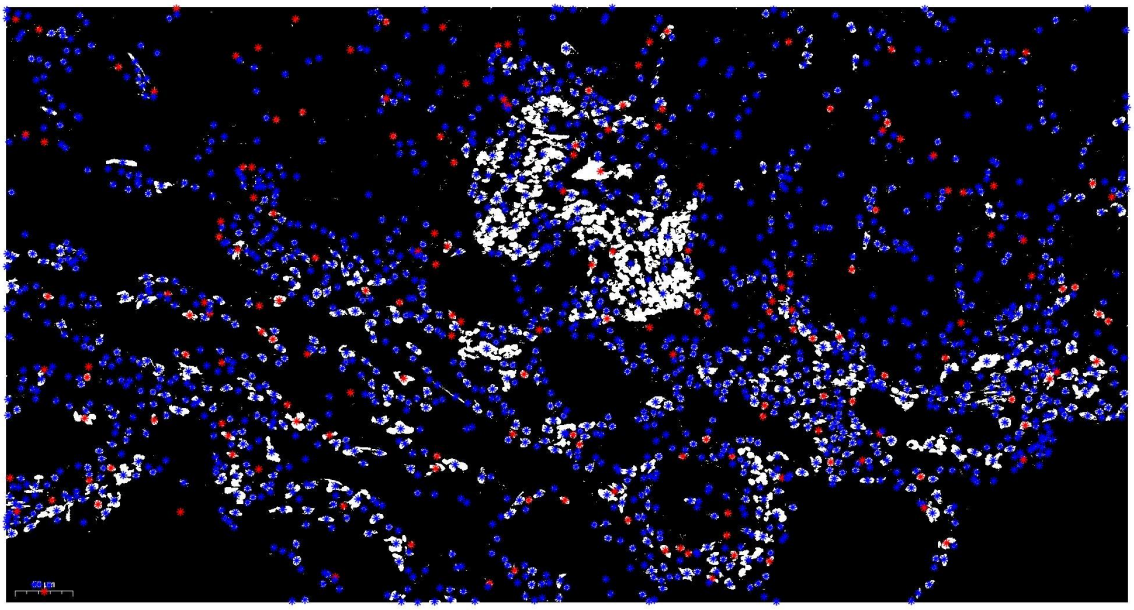

Figure 48. Metastatic tumor patient 86, blue cells have been classified correctly and red cells have been classified wrongly

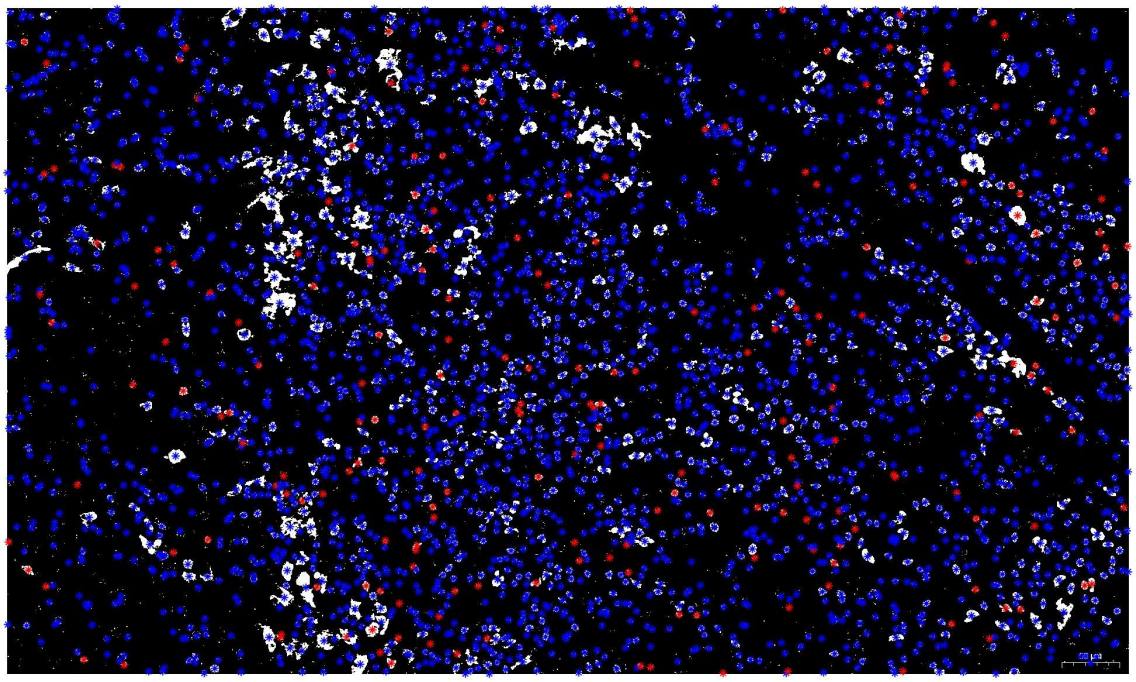

Figure 49. Metastatic tumor patient 105, blue cells have been classified correctly and red cells have been classified wrongly

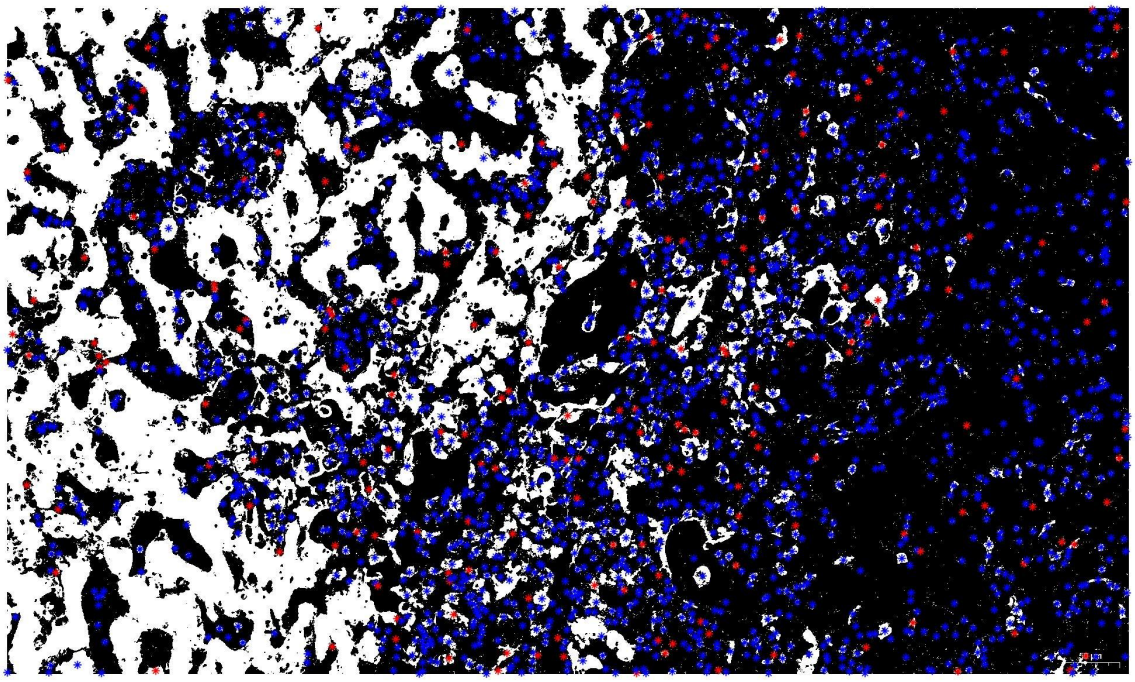

Figure 50. Metastatic tumor patient 8, blue cells have been classified correctly and red cells have been classified wrongly

4.5.2. Classification-Morphological features

In this model, morphological features of the cells were extracted to be used in the SVM classifier. Our desire was to see if primary and secondary tumor cells were different in shape. The features that are used here are as follows:

Area: Actual number of pixels in the object.

Major Axis Length and Minor Axis Length: Length of the major axis and minor axis of the ellipse, respectively.

Eccentricity: Eccentricity which its value is between zero and one is the measure of how nearly circular the ellipse is.

Solidity: area of an object divided by the area of its convex hull.

Perimeter: Distance around the boundary of an object.

Elongation: Elongation is the ratio between the length and width of the smallest rectangle containing the object in an image [70]-[71].

Several other morphology features such as circularity, extrema and etc. have been calculated and fed to the classifier to compare the results. The highest accuracy achieved in this method was 70%. It was observed that primary and secondary tumor cells are not that different in morphological aspects. This was expected, as the secondary cells are actually just primary cancer cells that break away from the primary tumor where they originated and travel via blood or lymph system to another organ to form a secondary tumor [72].

4.5.3. Classification- Antibody uptake in grayscale images

In this method, most cases (a case is defined, for example, when the classifier is trained based on the information of one patient and then tested with the information of another patient) achieved an accuracy of 90 % or higher. This method consisted of the normalized intensity related to the amount of protein expressions of different antibodies in primary and metastatic tumor cells being fed to the classifier.

The reason that accuracy in this method -- using normalized intensity of antibody uptake in grayscale images -- was much higher than the other methods is that primary and metastatic tumor cells are mostly the same in terms of morphology and antibody uptake,

but have an observable difference in their amount of protein expression. Immunofluorescence images are very sensitive and are therefore very suitable to be used to discern and compare how primary and secondary cells are different in the expression of their proteins. Even still, when the amount of protein is very low, the cells with low protein expression might be hidden behind the cells with high protein expression, in general with fluorescent images, the amount of signal in a cell has a linear relationship with the amount of protein expression in that cell.

Table 5 presents the information of the patients whose data has been used in the training and testing of the classifier. For instance, PM 80 in the Train column signifies that primary and metastatic tumor information of patient 80 has been used to train the classifier. Similarly, P116-M8 in the Test column means that the primary information of patient number 116 and the metastasis information of patient number 8 have been used to test the model. The 4 metrics, which include accuracy, precision, recall and f1 score, have been used to evaluate the performance of this classifier. The definition of these metrics and their formulas - with abbreviations TP, TN, FP and FN for True Positive, True Negative, False Positive and False Negative, respectively - are as follows: [73]

Accuracy: Number of correct predictions divided by total number of predictions

Accuracy = TP+TN/TP+FP+FN+TN

Precision: True positive divided by total number of true positive and false positive

Precision = TP/TP+FP

Recall: True positive divided by total number of true positive and false negative

Recall = TP/TP+FN

F1 score: 2 times precision multiply by recall divided by total number of precision and recall

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Though accuracy is very important in the assessment of the performance of a classifier, other metrics represent other significant aspects of that classifier as well. For instance, high precision implies a low false positive rate.

The highest accuracy in this method was 96 %. The other metrics confirmed the good performance of this classifier. Exceptions to this accuracy occurred for patients such as patient 116 for whom the Dapi stained scan of the cells was extremely low in intensity for the whole image and for patients for whom the metastasis information was not available in the dataset. In instances of the latter, the primary information of the patient had to be combined with the metastasis information of another patient in order to contribute to the classifier, therefore, the accuracy was not very high. Other than these two exceptions, the performance of the designed classifier was high. Nevertheless, this model should be examined in a larger dataset.

Train	Test	Accuracy	precision	recall	f1-score
PM 80	PM 105	96 %	0.72	1.00	0.84
PM 80	P116-M8	65%	0.67	0.96	0.79
PM 80	P105-ML8	91 %	0.91	1.00	0.95
PM 105	PM 80	94 %	1.00	0.88	0.94
PM 105	P116-M8	70%	0.78	0.77	0.78
PM 105	P80-M70	92 %	1.00	0.88	094
PM 105	P80-M10	93%	0.98	0.88	0.93
PM 105	P80-ML8	93%	1.00	0.88	0.94

Table 5. Metrics for different cases in the classification

As we noticed from the results, PCA could cluster the data into two different groups. in cases when the primary or metastatic data is both from the same patient, not different patients. Most of the variance is explained by PCs 1 and 3 -- the first and the third PCs. Using Vector Quantization, though, there is no discernable pattern to cluster the data to two or more different groups.

In supervised methods, we designed two types of classifiers; the first one was based on the morphological characteristics of primary versus metastatic tumor cells with the highest accuracy of 70%. We noticed that morphological features are not proper choices for distinguishing between primary and metastatic tumor cells, as the secondary cells are actually just primary cancer cells that break away from the primary tumor where they originated and travel via blood or lymph system to another organ to form a secondary tumor.

Our approach for designing the second type of the classifier was based on multiple antibodies uptake in the primary versus metastatic tumor cells. For this purpose, we first examined the presence of different antibodies in a cell and later the amount of protein expressions in different antibodies. We achieved the accuracy of 90% and higher in this method and the reason is that primary and metastatic tumor cells are mostly the same in

terms of morphology and antibody uptake, but have an observable difference in their amount of protein expression. Also, , in general with fluorescent images, the amount of signal in a cell has a linear relationship with the amount of protein expression in that cell and this makes this kind of image a good choice for detecting the amount of a specific antibody or a panel of different antibodies to be used in further analysis.

Chapter 5
Summary, Conclusions and, Future Work

The high mortality rate of pancreatic cancer and its poor response to common cancer treatments such as radiotherapy, immunotherapy, and chemotherapy have fueled research to develop novel therapeutic approaches for the diagnosis, treatment, and possibly cure of this highly lethal disease.

Tumor associated biomarkers are currently being intensively studied and have shown encouraging results for the management of pancreatic cancers. However, metastasis which is responsible for about 90 % of cancer deaths, is poorly understood. Understanding the biology and the dynamic of metastasis is crucial in the discovery and development of innovative therapies for this challenging cancer. Comparison of the absence or presence of validated biomarkers in primary versus secondary tumor sites can bring an insight into how the two tumors are different despite being originated from the same organ, and this can in turn help with revising the available or developing new treatment methods.

In this project, our ultimate goal was to design a classifier to classify pancreatic tumor cells into two categories; primary or metastatic. Most of the literature deals with the differences between normal cells and tumor cells and the classifiers are designed based on these differences. We aimed to examine the differences between primary versus metastatic tumor cells and investigate if the classifier could classify these cells based on two different criteria; morphological feature and the uptake of different antibodies.

For this purpose, after registering consecutive images of the same tissue slide together, first we investigated whether using unsupervised clustering methods such as Vector Quantization (VQ) and Principal Component Analysis (PCA) we could cluster primary metastatic cells into two different groups.

We noticed from the results, in cases when the data is both primary or metastatic but from different patients, PCA could not cluster the data into two different groups. However, when the dataset is primary-metastasis data from one patient, PCA has clustered the data to almost two different clusters. Nevertheless, the results overlap in some cases and this makes them not very reliable. Most of the variance is explained by PCs 1 and 3 -- the first and the third PCs.

In Vector Quantization, though, there is no specific pattern to distinguish between two or more different groups.

In the classification methods, we designed two types of classifiers; the first one was based on the morphological characteristics of primary versus metastatic tumor cells. In this method, features such as area, perimeter, solidity, eccentricity etc. were analyzed to examine if primary and metastatic tumor cells are morphologically different.

The highest accuracy achieved in this method was 70%. It was observed that primary and secondary tumor cells are not that different in morphological aspects. This was expected, as the secondary cells are actually just primary cancer cells that break away from the primary tumor where they originated and travel via blood or lymph system to another organ to form a secondary tumor.

The second type of the classifier was based on antibody uptake in the primary versus metastatic tumor cells. In this method, we first used the presence versus the absence of the antibodies in the binary images of our dataset as features to be fed to the classifier and later the normalized intensity related to the amount of protein expressions of different antibodies in primary and metastatic tumor cells.

In the first method, the highest accuracy reached was 82%, however, in the second method, most cases (a case is defined, for example, when the classifier is trained based on the information of one patient and then tested with the information of another patient) achieved an accuracy of 90 % or higher. This method consisted of the normalized intensity related to the amount of protein expressions of different antibodies in primary and metastatic tumor cells being fed to the classifier.

The reason that accuracy in this method -- using the normalized intensity of antibody uptake in grayscale images -- was much higher than the other methods is that primary and metastatic tumor cells are mostly the same in terms of morphology and antibody uptake, but have an observable difference in their amount of protein expression. Immunofluorescence whole slide images are highly sensitive and are therefore very suitable to be used to discern and compare how primary and secondary cells are different in the expression of their proteins. Even still, when the amount of protein is very low, the cells with low protein expression might be hidden behind the cells with high protein expression, in general with fluorescent images, the amount of signal in a cell has a linear relationship with the amount of protein expression in that cell.

Since the classifier is first trained based on the binary and later the grayscale images of the antibodies of primary and metastatic pancreatic cancer tissues, the next step in this work could be to investigate how the expression of a particular antibody or different panels of antibodies are quantitatively different in primary versus metastatic tumor cells. Also, since our dataset is considered a small dataset, further research will include examining the performance of the designed classifier on a larger dataset.

REFERENCES

[1] "The Pancreas | Johns Hopkins Medicine." [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/the-pancreas. [Accessed: 28-Mar-2021].

[2] M. Ilic and I. Ilic, "Epidemiology of pancreatic cancer," *World Journal of Gastroenterology*, vol. 22, no. 44. Baishideng Publishing Group Co., Limited, pp. 9694–9705, 2016, doi: 10.3748/wjg.v22.i44.9694.

[3] "Pancreatic Cancer: Statistics | Cancer.Net." [Online]. Available: https://www.cancer.net/cancer-types/pancreatic-cancer/statistics. [Accessed: 28-Mar-2021].

[4] "Biomarkers In Risk Assessment: Validity And Validation (EHC 222, 2001)." [Online]. Available: http://www.inchem.org/documents/ehc/ehc/ehc222.htm. [Accessed: 28-Mar-2021].

[5] M. Diamandis, N. M. A. White, and G. M. Yousef, "Personalized medicine: Marking a new epoch in cancer patient management," *Molecular Cancer Research*, vol. 8, no. 9. pp. 1175–1187, Sep-2010, doi: 10.1158/1541-7786.MCR-10-0264.

[6] L. Fass, "Imaging and cancer: A review," *Molecular Oncology*, vol. 2, no. 2. Wiley-Blackwell, pp. 115–152, Aug-2008, doi: 10.1016/j.molonc.2008.04.001.

[7] "Normal And Cancer Cells Structure: Image Details - NCI Visuals Online."

[Online]. Available: https://visualsonline.cancer.gov/details.cfm?imageid=2512. [Accessed: 28-Mar-2021].

[8] Y. H. Chang *et al.*, "Deep learning based Nucleus Classification in pancreas histological images," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, no. July, pp. 672–675, 2017, doi: 10.1109/EMBC.2017.8036914.

[9] I. Jansen *et al.*, "Histopathology: ditch the slides, because digital and 3D are on show," *World J. Urol.*, vol. 36, no. 4, pp. 549–555, Apr. 2018, doi: 10.1007/s00345-018-2202-1.

[10] "What is Digital Pathology? Benefits, Future Trends & More : Leica Biosystems." [Online]. Available: https://www.leicabiosystems.com/knowledge-pathway/digital-pathology/. [Accessed: 10-Nov-2020].

[11] M. Indu, R. Rathy, and M. P. Binu, "'Slide less pathology': Fairy tale or reality?," *Journal of Oral and Maxillofacial Pathology*, vol. 20, no. 2. Medknow Publications, pp. 284–288, 01-May-2016, doi: 10.4103/0973-029X.185921.

[12] M. D. Zarella *et al.*, "A practical guide to whole slide imaging a white paper from the digital pathology association," *Archives of Pathology and Laboratory Medicine*, vol. 143, no. 2. College of American Pathologists, pp. 222–234, 01-Feb-2019, doi: 10.5858/arpa.2018-0343-RA.

[13] "Brightfield Microscopy - Uses & Advancements; Microscope Reviews; Pros and Cons." [Online]. Available: https://www.microscopemaster.com/brightfield-microscopy.html. [Accessed: 28-Mar-2021].

[14] M. J. Sanderson, I. Smith, I. Parker, and M. D. Bootman, “Fluorescence microscopy,” *Cold Spring Harb. Protoc.*, vol. 2014, no. 10, pp. 1042–1065, Oct. 2014, doi: 10.1101/pdb.top071795.

[15] K. M. Gustashaw, P. Najmabadi, and S. J. Potts, “Measuring Protein Expression in Tissue,” *Lab. Med.*, vol. 41, no. 3, pp. 135–142, Mar. 2010, doi: 10.1309/LM8QU4SBNV2VSAST.

[16] “Multiplexed immunohistochemistry: Illuminating the tumor microenvironment to study cancer-immune mechanisms | Science | AAAS.” [Online]. Available: https://www.sciencemag.org/custom-publishing/webinars/multiplexed-immunohistochemistry-illuminating-tumor-microenvironment. [Accessed: 28-Mar-2021].

[17] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, “Digital Imaging in Pathology: Whole-Slide Imaging and Beyond,” doi: 10.1146/annurev-pathol-011811-120902.

[18] L. Pantanowitz, N. Farahani, and A. Parwani, “Whole slide imaging in pathology: advantages, limitations, and emerging perspectives,” *Pathol. Lab. Med. Int.*, vol. 7, p. 23, Jun. 2015, doi: 10.2147/PLMI.S59826.

[19] A. J. Evans *et al.*, “US Food and Drug Administration approval of whole slide imaging for primary diagnosis: A key milestone is reached and new questions are raised,” *Archives of Pathology and Laboratory Medicine*, vol. 142, no. 11. College of American Pathologists, pp. 1383–1387, 01-Nov-2018, doi: 10.5858/arpa.2017-

0496-CP.

[20] I. Girolami *et al.*, “Diagnostic concordance between whole slide imaging and conventional light microscopy in cytopathology: A systematic review,” *Cancer Cytopathol.*, vol. 128, no. 1, pp. 17–28, Jan. 2020, doi: 10.1002/cncy.22195.

[21] A. S. Azam *et al.*, “Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis,” *Journal of Clinical Pathology*, vol. 0. BMJ Publishing Group, pp. 1–8, 15-Sep-2020, doi: 10.1136/jclinpath-2020-206764.

[22] M. G. Hanna *et al.*, “Whole slide imaging equivalency and efficiency study: experience at a large academic center,” *Mod. Pathol.*, vol. 32, no. 7, pp. 916–928, Jul. 2019, doi: 10.1038/s41379-019-0205-0.

[23] F. Spill, D. S. Reynolds, R. D. Kamm, and M. H. Zaman, “Impact of the physical microenvironment on tumor progression and metastasis,” *Current Opinion in Biotechnology*, vol. 40. Elsevier Ltd, pp. 41–48, 01-Aug-2016, doi: 10.1016/j.copbio.2016.02.007.

[24] “Tumor Microenvironment - an overview | ScienceDirect Topics.” [Online]. Available: https://www.sciencedirect.com/topics/medicine-and-dentistry/tumor-microenvironment. [Accessed: 28-Mar-2021].

[25] R. Baghban *et al.*, “Tumor microenvironment complexity and therapeutic implications at a glance,” *Cell Communication and Signaling*, vol. 18, no. 1. BioMed Central Ltd., pp. 1–19, 07-Apr-2020, doi: 10.1186/s12964-020-0530-4.

[26] R. Wei, S. Liu, S. Zhang, L. Min, and S. Zhu, "Cellular and Extracellular Components in Tumor Microenvironment and Their Application in Early Diagnosis of Cancers," *Analytical Cellular Pathology*, vol. 2020. Hindawi Limited, 2020, doi: 10.1155/2020/6283796.

[27] "Role of the tumor microenvironment in pancreatic cancer," 2019, doi: 10.1002/ags3.12225.

[28] B. Ren *et al.*, "Tumor microenvironment participates in metastasis of pancreatic cancer," *Molecular Cancer*, vol. 17, no. 1. BioMed Central Ltd., pp. 1–15, 30-Jul-2018, doi: 10.1186/s12943-018-0858-1.

[29] P. J. Grippo and H. G. Munshi, *Pancreatic Cancer and Tumor Microenvironment*. Transworld Research Network, 2012.

[30] S. Wang *et al.*, "Tumor microenvironment in chemoresistance, metastasis and immunotherapy of pancreatic cancer," 2020.

[31] W. J. Ho, E. M. Jaffee, and L. Zheng, "The tumour microenvironment in pancreatic cancer — clinical challenges and opportunities," *Nature Reviews Clinical Oncology*, vol. 17, no. 9. Nature Research, pp. 527–540, 01-Sep-2020, doi: 10.1038/s41571-020-0363-5.

[32] J. Tao *et al.*, "Targeting hypoxic tumor microenvironment in pancreatic cancer," *Journal of Hematology and Oncology*, vol. 14, no. 1. BioMed Central Ltd, p. 14, 01-Dec-2021, doi: 10.1186/s13045-020-01030-w.

[33] E. Karamitopoulou, “The Tumor Microenvironment of Pancreatic Cancer,” *Cancers (Basel).*, vol. 12, no. 10, p. 3076, Oct. 2020, doi: 10.3390/cancers12103076.

[34] J. R. Lin *et al.*, “Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes,” *Elife*, vol. 7, Jul. 2018, doi: 10.7554/eLife.31657.

[35] R. Feldman and E. S. Kim, “Prognostic and predictive biomarkers post curative intent therapy,” *Annals of Translational Medicine*, vol. 5, no. 18. AME Publishing Company, 01-Sep-2017, doi: 10.21037/atm.2017.07.34.

[36] H. F. M. Kamel and H. S. B. Al-Amodi, “Cancer Biomarkers,” in *Role of Biomarkers in Medicine*, InTech, 2016.

[37] M. Herreros-Villanueva and L. Bujanda, “Non-invasive biomarkers in pancreatic cancer diagnosis: What we need versus what we have,” *Annals of Translational Medicine*, vol. 4, no. 7. AME Publishing Company, pp. 9–9, 01-Apr-2016, doi: 10.21037/atm.2016.03.44.

[38] J. S. Ankeny *et al.*, “Circulating tumour cells as a biomarker for diagnosis and staging in pancreatic cancer,” *Br. J. Cancer*, vol. 114, no. 12, pp. 1367–1375, Jun. 2016, doi: 10.1038/bjc.2016.121.

[39] J. Iovanna, “Implementing biological markers as a tool to guide clinical care of patients with pancreatic cancer,” *Translational Oncology*, vol. 14, no. 1. Neoplasia

Press, Inc., p. 100965, 01-Jan-2021, doi: 10.1016/j.tranon.2020.100965.

[40] S. Kato and K. Honda, “Use of Biomarkers and Imaging for Early Detection of Pancreatic Cancer,” *Cancers (Basel).*, vol. 12, no. 7, p. 1965, Jul. 2020, doi: 10.3390/cancers12071965.

[41] A. McGuigan, P. Kelly, R. C. Turkington, C. Jones, H. G. Coleman, and R. S. McCain, “Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes,” *World Journal of Gastroenterology*, vol. 24, no. 43. Baishideng Publishing Group Co., Limited, pp. 4846–4861, 21-Nov-2018, doi: 10.3748/wjg.v24.i43.4846.

[42] A. Litman-Zawadzka, M. Łukaszewicz-Zając, and B. Mroczko, “Novel potential biomarkers for pancreatic cancer – A systematic review,” *Advances in Medical Sciences*, vol. 64, no. 2. Medical University of Bialystok, pp. 252–257, 01-Sep-2019, doi: 10.1016/j.advms.2019.02.004.

[43] A. I. Kotzev and P. V. Draganov, “Carbohydrate Antigen 19-9, Carcinoembryonic Antigen, and Carbohydrate Antigen 72-4 in Gastric Cancer: Is the Old Band Still Playing?,” *Gastrointest. Tumors*, vol. 5, no. 1–2, pp. 1–13, 2018, doi: 10.1159/000488240.

[44] S. Hasan, R. Jacob, U. Manne, and R. Paluri, “Advances in pancreatic cancer biomarkers,” *Oncol. Rev.*, vol. 13, no. 1, pp. 69–76, 2019, doi: 10.4081/oncol.2019.410.

[45] N. S. Yee, S. Zhang, H. Z. He, and S. Y. Zheng, “Extracellular vesicles as potential biomarkers for early detection and diagnosis of pancreatic cancer,” *Biomedicines*, vol. 8, no. 12. MDPI AG, pp. 1–20, 01-Dec-2020, doi: 10.3390/biomedicines8120581.

[46] A. N. Ariston Gabriel *et al.*, “The involvement of exosomes in the diagnosis and treatment of pancreatic cancer,” *Molecular Cancer*, vol. 19, no. 1. BioMed Central Ltd, p. 132, 27-Aug-2020, doi: 10.1186/s12943-020-01245-y.

[47] Anuranjeeta, K. K. Shukla, A. Tiwari, and S. Sharma, “Classification of histopathological images of breast cancerous and non cancerous cells based on morphological features,” *Biomed. Pharmacol. J.*, vol. 10, no. 1, pp. 353–366, 2017, doi: 10.13005/bpj/1116.

[48] S. Stefanovic *et al.*, “Tumor biomarker conversion between primary and metastatic breast cancer: mRNA assessment and its concordance with immunohistochemistry,” *Oncotarget*, vol. 8, no. 31, pp. 51416–51428, 2017, doi: 10.18632/oncotarget.18006.

[49] D. S. Bhullar, J. Barriuso, S. Mullamitha, M. P. Saunders, S. T. O’Dwyer, and O. Aziz, “Biomarker concordance between primary colorectal cancer and its metastases,” *EBioMedicine*, vol. 40, pp. 363–374, 2019, doi: 10.1016/j.ebiom.2019.01.050.

[50] C. Gomez-Roca *et al.*, “Differential expression of biomarkers in primary non-small cell lung cancer and metastatic sites,” *J. Thorac. Oncol.*, vol. 4, no. 10, pp.

1212–1220, 2009, doi: 10.1097/JTO.0b013e3181b44321.

[51] D. Ansari, C. Urey, C. Gundewar, M. P. Bauden, and R. Andersson, “Comparison of MUC4 expression in primary pancreatic cancer and paired lymph node metastases,” *Scand. J. Gastroenterol.*, vol. 48, no. 10, pp. 1183–1187, Oct. 2013, doi: 10.3109/00365521.2013.832368.

[52] O. Déniz, D. Toomey, C. Conway, and G. Bueno, “Multi-stained whole slide image alignment in digital pathology,” in *Medical Imaging 2015: Digital Pathology*, 2015, vol. 9420, p. 94200Z, doi: 10.1117/12.2082256.

[53] J. B. A. Maintz and M. a Viergever, “An Overview of Medical Image Registration Methods (Cited by: 2654),” *Nature*, vol. 12, no. 6, pp. 1–22, 1996.

[54] B. Vidhyapeeth Rajasthan Neelam Sharma and A. Professor Banasthali Vidhyapeeth Rajasthan, “An Overview of Various Template Matching Methodologies in Image Processing,” *Int. J. Comput. Appl.*, vol. 153, no. 10, pp. 975–8887, 2016.

[55] S. Y. Chen, H. Qian, Z. Wu, and M. L. Zhu, “Fast normalized cross-correlation for template matching,” *Chinese J. Sensors Actuators*, vol. 20, no. 6, pp. 1325–1329, 2007.

[56] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-Up Robust Features (SURF),” *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008, doi: 10.1016/j.cviu.2007.09.014.

[57] E. Karami, S. Prasad, and M. Shehata, “Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images.”

[58] “Unsupervised Learning and Data Clustering | by Sanatan Mishra | Towards Data Science.” [Online]. Available: https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a. [Accessed: 29-Mar-2021].

[59] K. Sayood, *Introduction to Data Compression*, 5th ed. Morgan Kaufmann, 2017.

[60] Y. Linde, A. Buzo, and R. M. Gray, “An Algorithm for Vector Quantizer Design,” *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, 1980, doi: 10.1109/TCOM.1980.1094577.

[61] “A Step-by-Step Explanation of Principal Component Analysis.” [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis. [Accessed: 29-Mar-2021].

[62] J. Tang, S. Alelyani, and H. Liu, “Feature Selection for Classification: A Review.”

[63] “(Tutorial) Feature Selection in Python - DataCamp.” [Online]. Available: https://www.datacamp.com/community/tutorials/feature-selection-python. [Accessed: 29-Mar-2021].

[64] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.

[65] G. Liu, S. Mao, and J. H. Kim, “A mature-tomato detection algorithm using

machine learning and color analysis," *Sensors (Switzerland)*, vol. 19, no. 9, May 2019, doi: 10.3390/s19092023.

[66] "Finding Non-Linear Decision Boundary in SVM | by Sourodip Kundu | Medium." [Online]. Available: https://medium.com/@KunduSourodip/finding-non-linear-decision-boundary-in-svm-a89a97a006d2. [Accessed: 29-Mar-2021].

[67] "(No Title)." [Online]. Available: https://cw.fel.cvut.cz/b201/_media/courses/a6m33bio/otsu.pdf. [Accessed: 29-Mar-2021].

[68] "S. Beucher and C. Lantuejoul, 'Use of Watersheds in Contour Detection,' International Workshop on Image Processing Real-Time Edge and Motion Detection/Estimation, Rennes, 1979. - References - Scientific Research Publishing." [Online]. Available: https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55))/reference/ReferencesPapers.aspx?ReferenceID=986878. [Accessed: 29-Mar-2021].

[69] "Definition of metastasis - NCI Dictionary of Cancer Terms - National Cancer Institute." [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metastasis. [Accessed: 29-Mar-2021].

[70] "Major and Minor Axes of an Ellipse - Expii." [Online]. Available: https://www.expii.com/t/major-and-minor-axes-of-an-ellipse-5110. [Accessed: 29-Mar-2021].

[71] "(No Title)." [Online]. Available: https://pats.cs.cf.ac.uk/@archive_file?p=366&n=final&f=1-Corey_White_C1133127_Dissertation_Medical_Image_Processing_-_Lesions.pdf&SIG=07f000c062b2cfe38933e1335f29b2ed2e50e984adb95632ddbf5ce62f3bd036. [Accessed: 29-Mar-2021].

[72] "What is Metastasis? | Cancer.Net." [Online]. Available: https://www.cancer.net/navigating-cancer-care/cancer-basics/what-metastasis. [Accessed: 29-Mar-2021].

[73] "Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog." [Online]. Available: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/. [Accessed: 29-Mar-2021].